

Some pages of this thesis may have been removed for copyright restrictions.

If you have discovered material in Aston Research Explorer which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our [Takedown policy](#) and contact the service immediately (openaccess@aston.ac.uk)

Quantifying Uncertainty in Citizen Weather Data

Simon Joseph Bell

Doctor of Philosophy

ASTON UNIVERSITY

Submitted: November 2014

© Simon Joseph Bell, 2014

Simon Joseph Bell asserts his moral right to be identified as the author of this thesis.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without appropriate permission or acknowledgement.



Thesis Summary

Quantifying Uncertainty in Citizen Weather Data

Simon Joseph Bell

Doctor of Philosophy

The sheer volume of citizen weather data collected and uploaded to online data hubs is immense. However as with any citizen data it is difficult to assess the accuracy of the measurements. Within this project we quantify just how much data is available, where it comes from, the frequency at which it is collected, and the types of automatic weather stations being used. We also list the numerous possible sources of error and uncertainty within citizen weather observations before showing evidence of such effects in real data. A thorough intercomparison field study was conducted, testing popular models of citizen weather stations. From this study we were able to parameterise key sources of bias. Most significantly the project develops a complete quality control system through which citizen air temperature observations can be passed. The structure of this system was heavily informed by the results of the field study. Using a Bayesian framework the system learns and updates its estimates of the calibration and radiation-induced biases inherent to each station. We then show the benefit of correcting for these learnt biases over using the original uncorrected data. The system also attaches an uncertainty estimate to each observation, which would provide real world applications that choose to incorporate such observations with a measure on which they may base their confidence in the data. The system relies on interpolated temperature and radiation observations from neighbouring professional weather stations for which a Bayesian regression model is used. We recognise some of the assumptions and flaws of the developed system and suggest further work that needs to be done to bring it to an operational setting. Such a system will hopefully allow applications to leverage the additional value citizen weather data brings to longstanding professional observing networks.

Key Words

Amateur, Bias, User-contributed, Bayesian

Acknowledgements

This research is funded by an EPSRC CASE award (10002388) with the Met Office.

Thanks are due to:

Dan Cornford – For conceptualising and instigating such an interesting project, and for his priceless support, wisdom and patience from start to finish.

Lucy Bastin – For her fantastic guidance, GIS wizardry, and invaluable critique.

Mike Molyneux (Met Office) – For his support and expert advice throughout the project.

Shona Hogg & Fiona Carse (Met Office) – For providing Met Office MMS data.

Stephen Burt (RMetS Fellow) - For his help interpreting the field study results and for sharing his own VP2 field study data.

Duick Young (University of Birmingham) – For his help setting up the intercomparison field site and providing additional sensor data.

Richard Jones (Aston University), Jonathan Blower (BBC) & Richard Dean (IE Design) – For their web development expertise and advice.

Alan Hewitt (Met Office) – For continued support regarding many aspects of the project, especially accessing and interpreting Met Office radar data.

Paula Taylor and Gareth Dow (Met Office) – For their advice on accessing Met Office UKV data.

Bruce Ingleby and Katy Campbell (Met Office) – For their insights, especially during the project's infancy.

Table of Contents

1. Introduction	21
1.1. Thesis structure	23
1.2. Data structure	24
1.3. Publications	26
2. Citizen meteorology.....	27
2.1. The current state of citizen observations.....	27
2.1.1. Spatial resolution.....	29
2.1.2. Temporal resolution	31
2.1.3. Common automatic weather stations.....	32
2.1.4. CWS metadata.....	33
2.2. Applications of CWS data	37
2.2.1. Previous applications	37
2.2.2. Potential applications	37
2.2.3. Useful variables	38
2.2.4. A need to quantify uncertainty	39
2.3. Sources of uncertainty	40
2.3.1. Calibration issues.....	40
2.3.2. Design flaws	41
2.3.3. Communication and software errors	42
2.3.4. Metadata issues.....	43
2.3.5. Representativity error	45
2.4. Summary.....	47
3. Parameterising station bias	48
3.1. Field study design.....	48
3.1.1. The test site	48
3.1.2. Tested citizen weather stations.....	50
3.2. Investigation	54
3.2.1. Temperature	54
3.2.2. Parameterising temperature biases	62

3.2.3.	Relative humidity and dew point	65
3.2.4.	Rainfall	71
3.3.	Summary.....	75
4.	Interpolating professional observations.....	77
4.1.	Input MMS data	78
4.2.	Case study periods.....	79
4.2.1.	Autumn.....	79
4.2.2.	Winter.....	80
4.2.3.	Spring	80
4.2.4.	Summer	81
4.3.	Interpolation model design.....	81
4.3.1.	Linear regression.....	82
4.3.2.	Bayesian framework	83
4.3.3.	Alternative clustered approach	85
4.4.	Basis Functions	88
4.4.1.	Met Office short range forecast	91
4.4.2.	Easting, northing and elevation.....	94
4.4.3.	Coastality.....	95
4.4.4.	Urbanisation	96
4.4.5.	Radial basis functions.....	97
4.5.	Model performance.....	98
4.5.1.	Predictive power.....	112
4.6.	Summary.....	113
5.	Modelling citizen station bias	115
5.1.	Input CWS data	115
5.2.	Exploratory analysis of WOW and MMS data	116
5.3.	Addressing radiation bias	121
5.3.1.	Clear-sky global horizontal irradiance.....	122
5.3.2.	Satellite imagery	123
5.3.3.	Model structure.....	125

5.3.4.	Model performance	128
5.4.	Addressing station design	133
5.4.1.	Automatic extraction of model name from metadata	133
5.4.2.	Design classes.....	134
5.4.3.	Evidence of design effect in WOW data	137
5.5.	Addressing representativity	139
5.5.1.	Station classifier web application.....	140
5.5.2.	Exploratory analysis.....	143
5.6.	Model framework.....	147
5.6.1.	Concept	147
5.6.2.	Initial CWS quality control.....	151
5.6.3.	Bayesian update procedure	153
5.6.4.	Computational resources	161
5.7.	Model performance	162
5.7.1.	Corrected CWS vs. Interpolated MMS	163
5.7.2.	Interpolating with corrected CWS data	178
5.8.	Summary.....	182
6.	Conclusion.....	184
6.1.	Contributions	184
6.2.	Implementing operationally	185
6.3.	Advice.....	186
6.4.	Further work	188
7.	References.....	190
8.	Appendix.....	199
8.1.	Equation notation.....	199
8.2.	Weather Underground station types.....	205
8.3.	WOW station types	205
8.4.	WOW count code	206
8.5.	WOW site ratings scheme	208
8.6.	Winterbourne No. 2 field study site metadata	210

8.7.	CWS versus MMS dew point temperature	215
8.8.	Rain gauge lab test results	217

Table of Figures

<i>Figure 1.1. A selection of common CWS. These particular station models are tested against standard professional equipment in an intercomparison field study (Section 3).</i>	21
<i>Figure 1.2. Flow diagram illustrating data propagation and model interaction within the complete CWS quality control system.</i>	25
<i>Figure 2.1. A time series of the number of stations uploading data to WOW (Weather Observations Website; wow.metoffice.gov.uk) around midday each day from WOW's launch in spring 2011 through to summer 2014. The list of stations is extracted from the JSON formatted data structure used to render the observations on the WOW landing page. This process can be performed daily as a Cron Job as detailed in Appendix 8.4.</i>	27
<i>Figure 2.2. Spatial distribution of weather station networks over Great Britain on the 1st June 2014. The figure shows the professional Met Office MMS network alongside two popular citizen networks: WOW and Weather Underground. Met Office station locations taken from their, regularly updated, metadata files. WOW locations found by directly calling the application program interface (API) that provisions the front-end website. Weather Underground locations extracted from the website's HTML via web scraping techniques.</i>	29
<i>Figure 2.3. Atmospheric scale definitions, adapted from Thunis & Bornstein (1996). The mean minimum distance between stations has been added to express how station density correlates with the various scales of atmospheric phenomena.</i>	30
<i>Figure 2.4. Spatial distribution of weather station networks over the London conurbation on the 1st June 2014. The figure shows the professional Met Office MMS network amongst the citizen networks WOW and Weather Underground. The overlaid grid is the ~1.5 km grid used in Met Office's short-range forecast model (UKV). UKV grid coordinates were first converted from an equatorial lat/lon projection to WGS84 using an internal Met Office function, before reprojected in esri's ArcGIS software, along with the WGS84 formatted station locations, to the OSGB36 projection of the underlying LCM2007 land cover map (Morton, et al., 2011) shown.</i>	31
<i>Figure 2.5. Number of UK stations uploading temperature observations to a) WOW and b) Weather Underground at the specified upload frequencies on the 1st June 2014. The observation frequency was derived after having web scraped a day's worth of observations from every station that uploaded at least 1 observation that day to wow.metoffice.gov.uk and www.wunderground.com.</i>	32
<i>Figure 2.6. Weather station manufacturers and models used to automatically upload data to Weather Underground in February 2012. A total of 1353 stations were investigated, of which 16.6% were of unknown type and have been excluded from the diagram (Appendix 8.2). Davis 'Plus' models incorporate solar radiation and UV sensors, 'FARS' stands for Fan-Aspirated Radiation Shield.</i>	33

<i>Figure 2.7 Screenshot from wow.metoffice.gov.uk showing the rating system a citizen is encouraged to complete when registering a station on WOW. A list of possible values for these attributes is shown in Appendix 8.5.</i>	34
<i>Figure 2.8. Distribution of a) Star and b) Temperature ratings on the Met Office’s WOW website as rated by the owners of each station. Ratings of 5 and A respectively correspond to the highest standard. A full breakdown of these ratings is given in Appendix 8.5. To retrieve these ratings for every WOW users a web scraper was written to extract this metadata from the webpage of each station. Sample size = 1361 stations (includes non-UK stations).</i>	34
<i>Figure 2.9. Word clouds visually representing the textual metadata provided by WOW members in the a) Site Description & b) Additional Information sections used to describe their stations to other members. Produced using the online tool at www.wordle.net. The size of the word indicates how frequently it is used by the WOW community as a whole. In November 2013, when this information was extracting from WOW using web scraping techniques (Section 2.1), only 640 out of 1100 WOW stations had any Site Description text available, and only 363 had listed Additional Information.</i>	36
<i>Figure 2.10. The Fine Offset WH1080 rain gauge. Note its small rectangular shape (51 × 111 mm), and small tipping buckets.</i>	42
<i>Figure 2.11. Comparison of user-contributed station elevation against the height for that location extracted from the GMTED2010 digital elevation model (DEM; Section 4.4.2). Solid line indicates the 1:1 line. The dashed line shows the 1:1 line were the metadata height given in feet.</i>	44
<i>Figure 2.12. A conceptual representation of source areas (footprints). Image from WMO (2010), Chapter 11. The dark shaded ellipses show the theoretical source area for sensors responding to the turbulent transport such as thermometers, 50% or 90% of the signal comes from the area inside the respective ellipses. They are dynamic, moving with wind speed and direction, and atmospheric stability.</i>	46
<i>Figure 3.1. The Met Office’s Winterbourne No. 2 weather station. The site includes sensors operated by the Met Office, the University of Birmingham, and the seven CWS being tested as part of this study. Facing in a North-Easterly direction.</i>	50
<i>Figure 3.2. Mean temperature bias at different hours of the day (UTC) and months of the year for three of the CWS tested. (a) Davis VP2(1), (b) Davis Vue(2) and (c) La Crosse WS2350. Note the change in the colour scale for the final plot. The values written in grey are the mean bias of each cell. These values are simply the average of all bias values that fall within a given hour and month division. The VP2 and Vue sample every 10 minutes, whereas the WS2350 samples every hour.</i>	55
<i>Figure 3.3. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 26 May 2013. A time series of MMS global radiation is shown in orange.</i>	57

Figure 3.4. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 19 th Feb 2013. A time series of MMS global radiation is shown in orange.....	58
Figure 3.5. Temperature bias as a function of global radiation levels for the seven CWS tested. Global radiation observations less than 0 W m ⁻² were rounded up to 0 W m ⁻²	58
Figure 3.6. Relationship between global radiation and temperature bias for the a) Oregon Scientific WMR200 and b) La Crosse WS2350. The red and green lines show 1 st and 2 nd order regression models fitted to the data.	59
Figure 3.7. The WMR200 station's temperature bias as a function of wind speed and global radiation. The mean bias for a given radiation (wind speed) bin for all wind speeds (radiation levels) is shown along the bottom (right side). Here the number within each cell signifies the sample size.	60
Figure 3.8. Thermal images taken from the southwest by a Flir i5 thermal imaging camera on a sunny summer afternoon: (a) Stevenson screen and (b) VP2, (c) Vue, (d) WMR200, (e) WS2350 and (f) WH1080 stations. All stations were in direct sunlight, and had been for several hours. The colour-scale is consistent. The white (hot) part of the VP2 station evident in panel (b) is its black rain gauge. For help identifying the parts of each station, cross-reference with Figure 1.1, but be aware of the change in perspective.	61
Figure 3.9. Demonstration of correcting CWS' temperature bias using a multiple linear regression model. The figure shows a histogram of the WS2350 station's temperature bias before and after the correction, along with a scatter plot of a sample of its observations overlaid with a grid of the learnt model. The data was randomly split in half to form the training and test datasets using daytime data only, for the WS2350 this resulted in 2222 training points and 2221 test points.	62
Figure 3.10. Kernels used to weight preceding global radiation (at 1 minute resolution) measurements over a 60 minute window.	63
Figure 3.11. Correlation (Pearson's linear correlation coefficient) between each station's temperature bias and 60 minutes' worth of preceding global radiation observations (using 1 minute resolution radiation data) that have been weighted using a selection of different weighting kernels. Only observations during the day are used.	64
Figure 3.12. Relationship between temperature bias and 60 minutes worth of preceding global radiation observations (using 1 minute resolution radiation data) that have been weighted using a selection of different weighting kernels. The relationship is quantified using the R ² statistic from a 2 nd order regression model used to predict the temperature bias from weighted global radiation observations. Only observations during the day are used.	65
Figure 3.13. Time series of the Vue(1) station's relative-humidity bias, that is Vue(1) humidity – MMS humidity.....	66

Figure 3.14. Relative humidity time series. Covers the period when the MMS's Rotronic HygroClip was changed, as indicated by the red line. Note the addition of the University of Birmingham (UoB) Vaisala humidity observations.	67
Figure 3.15. The CWS' versus MMS's relative humidity: (a) VP2(1), (b) WMR200, (c) WS2350 and (d) WH1080 stations. All observations between the 16 th May 2013 and 31 st Aug 2013 are shown. The darker the colour the greater the density of points.	68
Figure 3.16. The CWS' versus MMS's dew-point temperature: (a) WMR200 and (b) WH1080 stations. All observations between the 16 th May 2013 and 31 st Aug 2013 are shown. The darker the colour the greater the density of points. The equivalent plots for the other CWS tested are show in Appendix 8.6.....	69
Figure 3.17. Plots of the relationship between temperature, humidity and dew point, and their biases for the La Crosse WS2350 station. All observations between the 16 th May 2013 and 31 st Aug 2013 are shown. The darker the colour the higher the density of points.	70
Figure 3.18. Cumulative rainfall totals of the seven CWS throughout the year-long study. Data from the professional Met Office gauge are shown by the black line. The final totals are displayed in the legend.....	72
Figure 3.19. Davis Vantage Vue on 24 th January 2013. Note that the rain gauge located on top of the unit is completely filled with snow, so much so that it prevents the wind cups from fully rotating.....	74
Figure 3.20. Plot of cumulative rainfall totals from the MMS's and seven CWS' gauges. The cumulative rainfall values (available every 10 min) for all CWS were corrected using the relationship between their cumulative rainfall and that of the MMS's gauge; learnt from preceding data only.....	75
Figure 4.1. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Autumn period: 1st – 14th October 2012. Synoptic charts from www.wetterzentrale.de archive..	79
Figure 4.2. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Winter period: 17th – 30th January 2013. Synoptic charts from www.wetterzentrale.de archive.....	80
Figure 4.3. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Spring period: 13th – 26th May 2013. Synoptic charts from www.wetterzentrale.de archive.	81
Figure 4.4. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Summer period: 24th June – 7th July 2013. Synoptic charts from www.wetterzentrale.de archive.	81
Figure 4.5. Screenshot of the graphical user interface (GUI) used to visualise the evolution of the clusters through time. The colour of each station represents the cluster to which its membership is strongest, shown here on the map (left) for a single timestep. The time series on the right of the plot shows the mean regression coefficient values for every basis function evolving over the 2 week summer period, at 3 hourly intervals. Each of the 3 clusters is represented by a different line/colour. Only MMS stations (225 total) were used here.	86

Figure 4.6. Screenshot of a GUI used to visualise the cluster assignments and how they correspond to the variables used to assign the clusters. The colour of each station in the Clusters plot denotes the cluster to which its membership is strongest (i.e. the cluster to which it has the greatest probability of belong to) and the size of each marker is indicative of the strength of its assignment to that cluster. Only MMS stations were used here.	87
Figure 4.7. Snapshots of the clusters evolving through time, here at 3-hourly intervals. Colour denotes strongest cluster membership for each station. Only MMS stations are shown here.	88
Figure 4.8. Time series of the sum of cluster weightings for each cluster, with each line representing a different cluster.	88
Figure 4.9. Met Office UKV T+3 forecast of 1.5 m air temperature field for 26 th June 2013 (summer period) at 03:00. Overlaid with MMS air temperature observations (circles) at the equivalent timestep.	91
Figure 4.10. UKV RMSE time series during the autumn period (1 - 14 October 2012) for different model cell height to station height temperature corrections; verified against observations from 225 MMS stations over the 112, 3-hourly, timesteps. Vertical grid lines mark midnight.	92
Figure 4.11. UKV bias arranged by 3-hourly timestep (rows) for each day (columns) during the summer period. Verified against observations from 207 MMS stations, of which 20 MMS stations were partially missing data, but were still used when data was available. Values are the mean bias of all the individual station biases at a given time when verified against MMS station observations. Bilinear interpolation is used to map predictions for the 4 nearest grid cells to station locations. A fixed lapse rate correction was used to adjust model predictions to MMS station heights.	94
Figure 4.12. Comparison of the MMS station elevations as listed within Met Office metadata with the GMTED2010 DEM derived elevation. The raster DEM was sampled at the station coordinates using bilinear interpolation.	95
Figure 4.13. a) Coastality and b) Urbanisation estimates across the British National Grid study domain (700 × 1300 1 km cells).	96
Figure 4.14. Distribution of RBF centres having use K-means to locate them over land areas only. Rings around each centre represent 1 standard deviation.	98
Figure 4.15. 1:1 plot of interpolation model temperature predictions against MMS observations using 10-fold cross-validation. Data shown is for all four 2 week periods; run separately, but plotted together. Magenta line is a 1:1 line. The darker the colour the higher the density of points.	99
Figure 4.16. Histogram of residuals when the interpolation model was verified against MMS observations using 10-fold cross-validation. The residuals for all four periods combined are shown.	100
Figure 4.17. Plot of z-scores for all four periods combined. Ideally ~95% of the z-scores should fall within the two red lines at ± 2 . Here 93.5% fall within this range. The z-score, $z = \frac{x - \mu}{\sigma}$, where x	

is the MMS observation, μ is the interpolation model's mean prediction, and σ is the estimated standard deviation of the prediction (i.e. the estimated uncertainty).	101
Figure 4.18. Rank histogram for all four periods combined. A flat, uniform, appearance implies a good probabilistic model (Hamill, 2001).	101
Figure 4.19. Coverage plot for all four periods combined. It plots the theoretical centred confidence interval against the observed frequency. When the points fall close to the red line the model has validated well probabilistically.	102
Figure 4.20. Reliability diagram for all four periods combined. This diagram compares forecast probabilities with actual observed frequencies (Bröcker & Smith, 2007), computed by splitting the range of observations into 10 classes. The number attached to each point denotes the number of observations in each class.	102
Figure 4.21. Time series of root mean squared error (RMSE) of the UKV model and the Interpolation model over each of the four 2-week case study periods. Both models were verified against the same set of 3-hourly air temperature observations from the ~220 MMS stations. Below each plot is a time series of the standard deviation of the observations. Vertical grid lines represent midnight at the start of each date.	105
Figure 4.22. Interpolation model RMSE of predictions made at the location of each MMS station. RMSE is calculated from the residuals of every timestep within the given period.	106
Figure 4.23. Spatial plots for the case study timestep of 10 th October 2012 06:00 (autumn period). The figure plots the MMS observations, the UKV model's predictions (height correction applied), the prediction from the interpolation model and its bias when verified against the MMS observations using 10-fold cross-validation. Note that the bottom right plot shows temperature bias rather than absolute temperature, and thus the colour scale has changed to accommodate for this.	108
Figure 4.24. Interpolation model error statistics when verified against each individual MMS stations' observations using cross-validation. Results here are for the summer period only. The order of the stations is based upon their elevation. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	109
Figure 4.25. Time series of the interpolation model cross-validation RMSE during the winter period, both when the regression coefficients are allowed to propagate through time, and when they are forgotten after each timestep and then re-learned from scratch at the next timestep.	110
Figure 4.26. Time series of the regression coefficient mean for the UKV basis function over the summer period. Each line represents a different fold within the 10-fold cross-validation.	111
Figure 4.27. Relationship between the UKV and interpolation model discrepancy (as quantified by the Hellinger distance) and the interpolation model error over the autumn period. The closer the Hellinger distance is to 0 the greater the agreement between the two models.	112

Figure 4.28. Each time series represents the change in RMSE when the given predictor is removed from the interpolation model; in this case for the summer period. A positive RMSE change implies the model accuracy decreases when the given predictor is left out.	113
Figure 5.1. Boxplots of difference between the uncorrected CWS observations and IMMS; plotted individually for each CWS station over the summer period. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	117
Figure 5.2. Visualisation of the difference between the uncorrected CWS observations and IMMS for each station (rows) and at each timestep (columns) over the summer period. Ticks on the x-axis indicate midnight at the start of that date.	118
Figure 5.3. Visible satellite images, from the Meteosat Second Generation satellite, during the summer period for the dates: a) 28 th June 2013 12:00, b) 6 th July 2013, 12:00. Source: BADC (badc.nerc.ac.uk).	119
Figure 5.4. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed star rating (higher = better). Values at the bottom denote the standard deviation. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	120
Figure 5.5. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed temperature rating. Ratings range from A, the highest quality, though to D the lowest. U denotes unknown quality and 0 implies a site with supposedly no temperature observations. Values at the bottom denote the standard deviation. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	120
Figure 5.6. Spatial distribution of Met Office MMS stations that regularly record Global Horizontal Irradiation (GHI).	121
Figure 5.7. Comparison of the clear-sky GHI estimates from two different approaches over the 1 st to 2 nd of June 2004.	123
Figure 5.8. Visible MSG satellite image of Great Britain on 24 th May 2013 at 14:00 GMT before (left) and after (right) removing the green country outlines and reprojecting to the British national grid.	124
Figure 5.9. 1:1 plots of radiation interpolation model predictions verified against MMS global radiation observations using 10-fold cross-validation. In a) point observations are used, whereas	

in b) observations over the past hour have been exponentially weighted. Only data from the summer period is shown. The deeper the colour the higher the density of points.....	126
Figure 5.10. Histogram of residuals from predictions made by the radiation interpolation model when using exponentially weighted global radiation observations as the target, verified against MMS observations using 10-fold cross-validation. Only data from the summer period is shown. The red line represents a Gaussian distribution fitted to the residuals.	127
Figure 5.11. 1:1 plot of predictions from the final radiation interpolation model against MMS observations when verified using 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.	129
Figure 5.12. Residuals from the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The Model includes both visible and infrared satellite imagery.....	129
Figure 5.13 Z-scores plot for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.....	130
Figure 5.14. Coverage plot for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery. It plots the theoretical centred confidence interval against the observed frequency.	131
Figure 5.15. Rank Histogram (Hamill, 2001) for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.....	131
Figure 5.16. 1:1 plot of predictions from the radiation interpolation model against MMS observations when verified with 10-fold cross-validation. All four periods are included in this plot. The model includes infrared satellite imagery, but not visible.	132
Figure 5.17. Residuals from the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes infrared satellite imagery, but not visible.....	132
Figure 5.18. Relationship between global radiation and the temperature bias for each of the 7 design classes as learnt from the intercomparison field study. The temperature bias is estimated based upon the equivalent lRad value for global radiation values of 0 W m^{-2} through to 1250 W m^{-2} at 1 W m^{-2} intervals. A second order regression function is used to predict the temperature bias from the equivalent lRad values using the regression coefficient mean terms $\mu\beta$ learnt from field study for CWS belonging to the given class.	136
Figure 5.19. Relationship between La Crosse WS2350 temperature bias and radiation (log transformed to lRad). Green line indicates the fitted model used to represent stations belonging to the 'Extreme Encased' design class.	136

Figure 5.20. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed model name and therefore designated design class. All available WOW stations (604 stations) are including using 3 hourly data throughout the summer period. Values at the bottom denote the standard deviation. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	138
Figure 5.21. Relationship between the temperature difference (uncorrected CWS – IMMS) and lRad for stations whose metadata assigns them to the design class a) Quality Louvered and b) Encased Louvered. The green line represents the temperature bias we would expect to for this design type, learnt from the intercomparison field study (Figure 5.18).	139
Figure 5.22. Screenshot of the options page for the Station Classifier web app.	141
Figure 5.23. Screenshot of the Station Classifier web application being used to classify WOW stations by exposure. Note how the user is dragging an aerial image to the column they feel is most appropriate.	142
Figure 5.24. Distribution of MMS stations into the various exposure (columns) and Urban Climate Zone (rows) classes. The values represent the number of stations, with darker colours indicating a higher count.	143
Figure 5.25. Distribution of CWS stations into the various exposure (columns) and Urban Climate Zone (rows) classes. The values represent the number of stations, with darker colours indicating a higher count. Every station counted submitted at least one observation during the summer period.	144
Figure 5.26. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their exposure classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.	145
Figure 5.27. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their UCZ classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.	145
Figure 5.28. Boxplots of the temperature difference between corrected CWS observations and IMMS after the CWS stations have been separated by their exposure classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.	146

Figure 5.29. Boxplots of the temperature difference between corrected CWS observations and IMMS after the CWS stations have been separated by their UCZ classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.....	146
Figure 5.30. Example of the learnt biases, modelled as Gaussian distributions, being subtracted from the raw CWS observation in order to correct it.	148
Figure 5.31. Theoretical time series over 5 days showing the CWS observation before and after a correction has been applied. The numbers correspond to the numbered steps below, which detail what is occurring at each step.....	150
Figure 5.32. Schematic of data flow through the bias correction model. Bracketed numbers refer to equation numbers in Section 5.6.3.	155
Figure 5.33. Visual representation of the overlap between the Predicted Radiation Bias for a given design class (Orange curve; e.g. Encased) and the estimated Observed Radiation Bias (Blue curve). The size of green overlap area is proportional to the scaling factor S	159
Figure 5.34. 1:1 plots of the interpolated MMS temperature observation against the a) uncorrected, and b) corrected CWS observations. Points shown are for all four 2 week case study periods. The magenta line indicates the 1:1 line.....	164
Figure 5.35. Box plot of the temperature discrepancy (CWS-IMMS) statistics for each individual station after the correction has been applied for the summer period. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red. Compare with (Figure 5.1), the equivalent figure for the uncorrected CWS data.....	165
Figure 5.36. Visualisation of the difference between the corrected CWS observations and IMMS for each station (rows) and at each timestep (columns) over the summer period. Ticks on the x-axis indicate midnight at the start of that date. Compare with Figure 5.2, the equivalent figure for the uncorrected CWS observations.	166
Figure 5.37. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias when the bias correction model was subjected to artificial CWS data with a calibration bias of $-2\text{ }^{\circ}\text{C}$ and radiation bias of $+0.007\text{ }^{\circ}\text{C per W m}^{-2}$. Red and blue shaded areas represent the uncertainty ($\pm 1\text{ s.d.}$) of the bias estimates.....	167
Figure 5.38. Change in design membership probabilities when the bias correction model was subjected to artificial CWS data with a calibration bias of $-2\text{ }^{\circ}\text{C}$ and radiation bias of $+0.007\text{ }^{\circ}\text{C per W m}^{-2}$. Initially the station model was unknown therefore each class was given an equal weighting.....	168
Figure 5.39. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 1. Red and blue shaded areas represent the uncertainty ($\pm 1\text{ s.d.}$) of the bias estimates.	169

Figure 5.40. Change in design membership probabilities for case-study Station 1. Initially the station model was unknown therefore each class was given an equal weighting.	169
Figure 5.41. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 2. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.	170
Figure 5.42. Change in design membership probabilities for case-study Station 2. Initially the design class was set as Encased Louvered based upon the station's metadata.	171
Figure 5.43. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 3. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.	172
Figure 5.44. Change in design membership probabilities for case-study Station 3. Initially the design class was set as Quality Louvered based upon the station's metadata.	172
Figure 5.45. Learnt observational uncertainty for each CWS station (rows) at each time (columns) during the summer period. Shown as the standard deviation, not the variance, i.e. $vs, tCWSc$	173
Figure 5.46. Learnt observational uncertainty for each CWS station (rows) at each time (columns) during the winter period. Shown as the standard deviation, i.e. $vs, tCWSc$	174
Figure 5.47. Distribution of the learnt calibration bias mean terms at the final timestep of the a) winter and b) summer periods.	174
Figure 5.48. Learnt calibration mean, $\mu_s, tCal$, and representativity term, shown as the standard deviation, i.e. $vs, tCWSc$, at the end of the summer period plotted spatially.	175
Figure 5.49. Number of stations allocated to each design class at the start (columns) vs the number at the end (rows) of the summer period. At the start the design class allocation were based upon the user's metadata.	176
Figure 5.50. Number of stations allocated to each design class at the start (columns) vs the number at the end (rows) over the summer period. Unlike Figure 5.50 each station was allocated an equal probability to each design class at the start. The metadata classes are still shown to assess whether the learnt classes at the end match the metadata.	177
Figure 5.51. Time series of cross-validation RMSE of the temperature interpolation model run under 3 scenarios: with MMS data only, with MMS data and uncorrected CWS data, with MMS data and corrected CWS data. Shown for each two week case study period: a) Autumn, b) Winter, c) Spring, d) Summer.	181
Figure 5.52. Coverage plot for when the temperature interpolation model was run with MMS data and corrected CWS data over the summer period. Verified against withheld MMS station observations using 10-fold cross-validation. It plots the theoretical centred confidence interval against the observed frequency.	182

Table of Tables

<i>Table 1. Summary of the 7 CWS tested as part of this field study.</i>	<i>52</i>
<i>Table 2. Key statistics from the field study</i>	<i>53</i>
<i>Table 3. Error statistics for the interpolation model during each of the four 2 week case study periods. Verified against MMS observations using 10-fold cross-validation.</i>	<i>103</i>
<i>Table 4. A list of the 7 design classes to which a CWS can be allocated based upon the style of radiation shielding and the subsequent radiation-induced biases exhibited.</i>	<i>135</i>
<i>Table 5. Mean and variance statistics for the discrepancy between CWS temperature observations and IMMS both before and after bias correction.</i>	<i>164</i>

Abbreviations

The following abbreviations are used throughout the thesis, please refer to this section as required.

CWS – A *Citizen Weather Station*. A weather station, which is usually low in cost, owned and setup by a citizen observer.

MMS – A professional Met Office land surface weather station, part of the network referred to as the *Meteorological Monitoring System*.

IMMS – *Interpolated MMS* temperature observations. Interpolated to the locations of the Citizen Weather Stations.

UKV – The Met Office’s short range, high resolution, numerical weather prediction model. It is a ~1.5 km resolution, ‘convection-permitting’, configuration of its Unified Model which covers the UK, hence its title of UKV.

GHI – *Global Horizontal Irradiance*. A measure of incoming solar irradiance (W m^{-2}) at a given surface location. Measures both direct and diffuse radiation combined.

UCZ – *Urban Climate Zone*. Areal land-cover zones classified by their capacity to modify the local climate (Oke, 2004).

RMSE – *Root Mean Square Error*. Used commonly herein as a measure of model error when model predictions are verified against observations.

RBFs – *Radial Basis Functions*. Used to provide localisation in our air temperature interpolation model. Only Gaussian RBFs are used herein.

API – *Application Program Interface*. Helps expose a program’s internal functions to other applications in a limited fashion. They allow information to be easily moved between different applications.

NWP – *Numerical Weather Prediction* model. Processes meteorological observations with computer models to forecast the future state of the weather.

1. Introduction

The growth of citizen science is evident within a wide range of scientific disciplines (Gura, 2013). Ignoring the data collected by these citizens would be a waste of often valuable additional information which is capable of increasing the spatial and temporal resolution of observing networks, whilst at the same time stimulating public engagement (Nov, et al., 2014). Crucially, citizen observation can do this at a fraction of the cost of professional systems.

Here we focus specifically on the field of meteorology, and in particular on meteorological observations collected automatically by low-cost, electronic, Citizen Weather Stations, referred to from now on as CWS (Figure 1.1). As many CWS submit their observations automatically to online data hubs such as the Met Office's Weather Observations Website (WOW – wow.metoffice.gov.uk) and Weather Underground (www.wunderground.com) the data is freely available for use in various applications and research projects, such as this study. As part of this project we review previous applications of this data and suggest possible future uses.

Very few studies have assessed the availability and characteristics of CWS data in the UK, with Bell, et al., (2013), Muller, et al., (2015) and Morris & Endfield (2012) providing notable exceptions. We therefore begin by using web scraping techniques to process data from these online hubs, allowing us to summarise the volume of information available. We show that with over 1800 CWS in the UK alone, this citizen network has the potential to add value to longstanding professional observing networks. By 'professional' we mean national meteorological organisations who abide by WMO standards (WMO, 2010). With a denser network of observations comes the possibility that fewer weather phenomena would go unobserved.



Figure 1.1. A selection of common CWS. These particular station models are tested against standard professional equipment in an intercomparison field study (Section 3).

However, as with many fields of citizen science, the data is prone to biases and increased uncertainty (Hunter, et al., 2013). By analysing data extracted from the online data hubs and by conducting our own empirical field study we were able to quantify the magnitude of these errors. Our year-long field study tested 5 unique brands of weather station (Figure 1.1) against collocated professional Met Office equipment. This is the first time such a numerous selection of CWS have been tested together at a single location. Previously studies have only collocated 1 or 2 types of CWS (Burt (2009); Burt (2013); Jenkins (2014)). We focus on measurements of temperature, relative humidity & dew point, and precipitation. The methodology and results of this intercomparison are detailed herein. Given the magnitude of the errors detected we strongly recommend a quality control procedure when handling CWS data.

We also present solutions for parameterising the bias in the CWS observations, using data from the field study to inform our methodology. We detail how CWS observations of air temperature suffer primarily from 2 key sources of bias. Firstly, and most significantly, were biases with a strong dependency on the strength of incoming solar radiation. The design of the station was seen to influence this relationship. Secondly calibration biases were recorded, which tend to remain relatively stable through time.

Having demonstrated that biases in CWS data are far from negligible our primary objective was to develop a mathematical model capable of correcting for these biases. This is not the first time a system has been developed to perform quality control checks on citizen data. The citizen weather observer program (CWOP; wxqa.com), for example, uses NOAA's Meteorological Assimilation Data Ingest System (madis.noaa.gov) to apply buddy-system quality control checks to the citizen data it receives. Nearby stations are used to model the weather at citizen locations, then the difference from the citizen observations is used to flag stations with poor internal, spatial, or temporal consistency. This strategy provides a very 'black or white' solution; either a station is flagged as erroneous or not. Lussana, et al., (2010) also developed a test to quality control temperature observations from an automatic meteorological network. Like CWOP, they aimed to identify gross errors and flag observations with poor spatial consistency. The online data hub Weather Underground also runs a quality control procedure on the CWS data it receives. Its operational *BestForecast* system leverages data it believes to be accurate to produce a site-specific forecast for the given CWS location.

In this thesis, we develop an alternative solution. A Bayesian framework is created capable of modelling bias in air temperature observations explicitly so that, instead of simply blacklisting observations, our approach can correct for any inherent biases. Crucially it does so whilst quantifying the uncertainty associated with the corrected observation. It is then up to users of the data to assess which observations are suitable for their own application based on the assigned, and validated, uncertainty estimates. This system directly models the 2 key sources of bias detected in the field study, i.e. the calibration bias, and radiation-induced bias. With this approach the quality of CWS across the UK can be assessed without having to physically visit each site in person. We present results showing the performance of this bias correction model, having tested the methodology on real citizen data from WOW.

A requirement of our bias correction model is that we have estimates of the air temperature and the strength of incoming solar radiation at the location of the CWS. We therefore demonstrate an approach for interpolating professional observations of these two variables, recorded at weather stations belonging to the UK Met Office's Meteorological Monitoring System (MMS). We demonstrate the benefit of using a Bayesian linear regression model to perform this interpolation, in particular showcasing its ability to quantify the uncertainty of the interpolated value, which is then propagated into the bias correction model.

We recognise that further work is needed to make such an approach operational and as such detail the challenges that must first be overcome to implement the approach demonstrated here operationally.

1.1. Thesis structure

Chapter 2 examines what CWS data is available; specifically, where the stations are, when they record, what stations are being used, and how they share their data. We go on to highlight the volume of data available before detailing previous projects that have already successfully used such citizen observations. This chapter concludes by introducing the possible sources of error and bias inherent to CWS measurements.

Having identified the likely sources of bias and uncertainty in CWS data, the logical next step was to collect our own CWS data to quantify the magnitude of those errors ourselves. This was carried out in the form of an intercomparison field study as detailed in Chapter 3, which tested common CWS against professional instruments. The focus was on measurements of temperature, relative humidity & dew point, and precipitation. For each we attempt to identify, explain and parameterise any biases.

Identifying and parameterising bias when professional instruments are collocated with the CWS is relatively straightforward, but in reality CWS sites are frequently tens of kilometres away from the nearest professional site. Therefore, an interpolation model was developed (Chapter 4) to interpolate professional temperature observations to CWS station locations. This provides an independent best estimate of the weather at the CWS locations, with an estimated uncertainty, against which the CWS observations can be verified.

It is then the job of the bias correction model, discussed in Chapter 5, to learn and correct any biases present within the data leaving only natural spatial variations in the temperature field. The process of distinguishing natural variations from artificial bias is informed by what was learnt in Chapter 3. We also explain how the bias correction model quantifies our confidence in the quality of the CWS data and any bias corrections we apply using associated uncertainty estimates.

Chapter 6 concludes by summarising this project’s key contributions whilst detailing the further work still required. It also offers advice to citizen observers, station manufacturers, data hubs and data users based upon the results of our investigation. It even details how such an approach could be implemented operationally and the challenges of doing so.

1.2. Data structure

This project combines many different sources of data which are passed through a variety of pre-processing functions, web applications and numerical models. Figure 1.1 provides an overview of how data propagates through our complete quality control system. Note how the final output is simply the CWS data which was originally fed into the system, but which has now been corrected for gross errors, calibration biases, and radiation-induced biases, and supplemented with an estimate of the uncertainty for each observation.

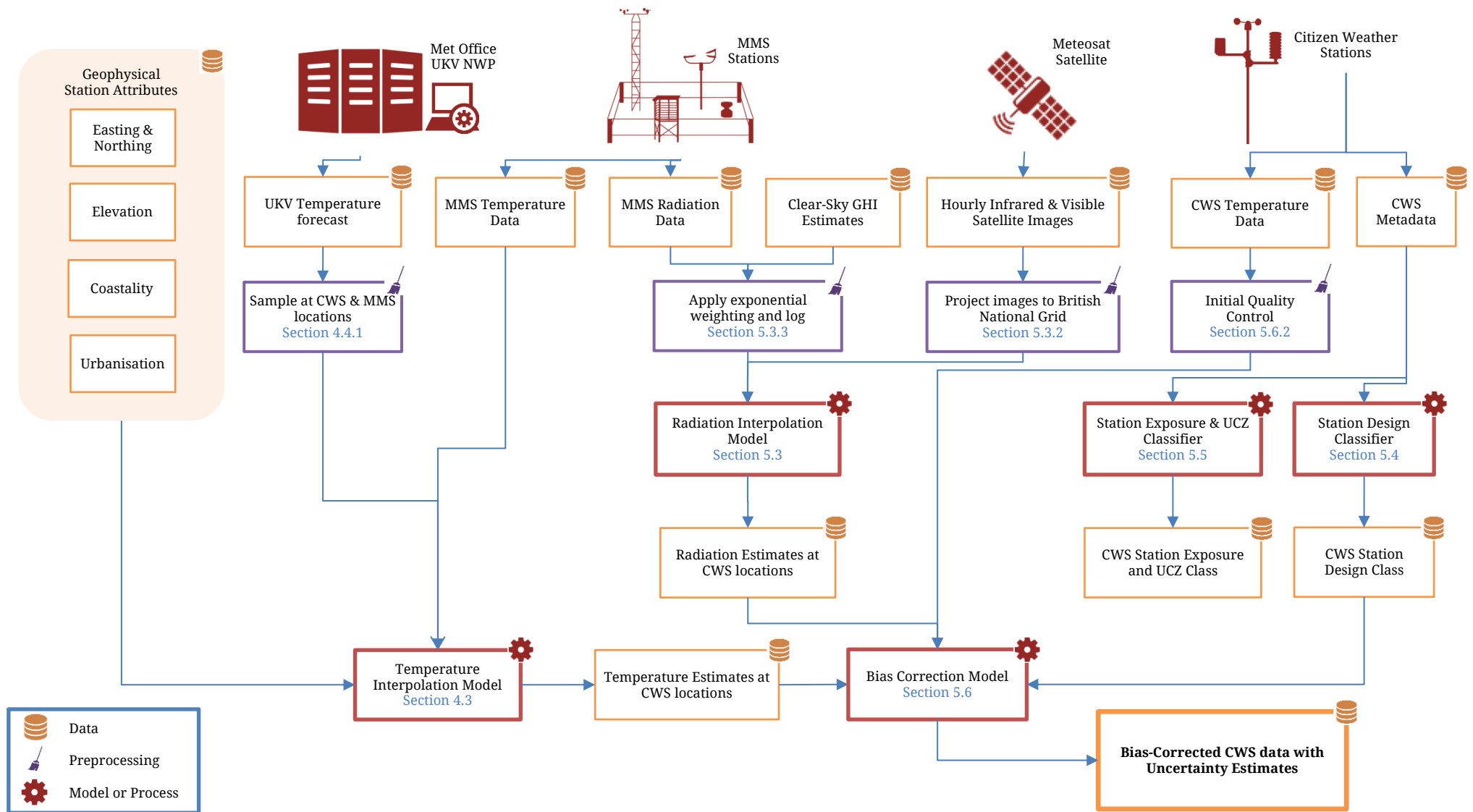


Figure 1.2. Flow diagram illustrating data propagation and model interaction within the complete CWS quality control system.

1.3. Publications

Bell, S., Cornford, D. & Bastin, L., 2013. The state of automated amateur weather observations. *Weather*, Volume 68(2), pp. 36-41. (Shares some content with Chapter 2)

Bell, S., Cornford, D. & Bastin, L. 2015. How good are citizen weather stations? Addressing a biased opinion. *Weather*. Volume 70(3), pp 75-84. (Shares some content with Chapter 3)

2. Citizen meteorology

In this thesis, we define a CWS as a weather station set up by a member of the public for whom the terms ‘weather enthusiast’, ‘volunteer’, ‘hobbyist’ and ‘amateur observer’ are fitting descriptions. Crucially, these stations are set up out of personal interest (or, in schools, for educational purposes) rather than because it is the owner’s job. Here we are particularly interested in automatic weather stations, which once installed can measure the state of the atmosphere at frequent intervals for their location with very little manual input. In this chapter we detail the where, when, what and how of citizen weather observing (Section 2.1), discuss what applications CWS data has been and could be used in (Section 2.2), and finally highlight the potential sources of uncertainty within CWS data (Section 2.3).

2.1. The current state of citizen observations

Citizen meteorology has been around for centuries: indeed, meteorology began thanks to the interest of amateurs (Eden, 2009). Currently over 1800 CWS are observing and recording the weather across the UK. For comparison, the Met Office runs 250 or so land-surface stations in its professional Meteorological Monitoring System (MMS) (Green, 2010).



Figure 2.1. A time series of the number of stations uploading data to WOW (Weather Observations Website; wow.metoffice.gov.uk) around midday each day from WOW’s launch in spring 2011 through to summer 2014. The list of stations is extracted from the JSON formatted data structure used to render the observations on the WOW landing page. This process can be performed daily as a Cron Job as detailed in Appendix 8.4.

This increase in numbers (Figure 2.1) is primarily due to the recent mass production of affordable and user-friendly weather stations. Such stations enable citizen meteorologists to automatically record sub-hourly observations (Figure 2.5), which can be stored electronically; allowing easy analysis and data sharing. Thus, manual observations at standard observing times, such as those taken by many members of the Climatological Observers Link (COL), are now supplemented with automated readings supplied by a wide range of observers.

Websites such as Weather Underground (www.wunderground.com, CWS hub launched in 2004) and the Met Office's more recently launched Weather Observations Website (WOW – wow.metoffice.gov.uk, launched in spring 2011) provide hubs to share the data with the wider community. These hubs allow citizens to access not only historical data but also near real-time observations anywhere in the world, and to compare their data with other citizen stations nearby. During the 24 hours that made up 1st July 2014, a total of 942 UK stations uploaded live data to the WOW website, with an equivalent 1870 stations uploading to Weather Underground. Under the assumption that a station on WOW within 100 m of a station on Weather Underground is the same station then 527 of these stations upload to both websites. The 100 m distance was selected to account for discrepancies in the station's location that may arise from a citizen having to use the two website's different interfaces to mark their station's location. Setting this distance any larger would increase the risk of separate stations on the same street being falsely classified as a single site.

Throughout this Section (2.1) we present data extracted from these two online data hubs. A process called *web scraping* was required to extract the relevant data from these two websites. Both websites serve a station's observations and its metadata within webpages written in HyperText Markup Language (HTML). By writing our own code in the programming language *Ruby*, with the aid of the *Ruby* gem *Nokogiri*, it was possible to extract the required data from the HTML. Such a web scraping approach was necessary as although WOW provides an application program interface (API) for citizens to submit their data, an equivalent API is not openly available to query observations already in the WOW data archive. Weather Underground does provide such an API, but in many instances the relevant data was not accessible via the API, in which case web scraping techniques were implemented. When accessing data this way it is important that the number of webpage requests per minute is kept low to prevent unnecessary strain on the servers that serve the webpages.

2.1.1. Spatial resolution



Figure 2.2. Spatial distribution of weather station networks over Great Britain on the 1st June 2014. The figure shows the professional Met Office MMS network alongside two popular citizen networks: WOW and Weather Underground. Met Office station locations taken from their, regularly updated, metadata files. WOW locations found by directly calling the application program interface (API) that provisions the front-end website. Weather Underground locations extracted from the website's HTML via web scraping techniques.

A significant attraction of CWS data is its spatial resolution. Figure 2.2 illustrates this point. The Met Office's purpose-designed observing network comprises of just over 200 observing sites (Green, 2010), the distribution of which is designed for an even coverage so that as few weather features as possible escape detection. The spread of citizen stations on the other hand is, unsurprisingly, clustered around major conurbations where their spatial density far exceeds the Met Office's network.

The *Near* tool from esri's ArcGIS software was used to calculate the mean minimum distance between the stations show in Figure 2.2, providing a measure of station density. For MMS stations only, the calculated distance is 19.2 km. By adding CWS stations from WOW and Weather Underground this distance falls to 4.2 km. In both cases any stations believed to be duplicates (i.e. < 100 m from the nearest site were removed). Figure 2.3 illustrates how these distances correspond to the scale of atmospheric phenomena. With CWS stations included, the network stands a greater chance of resolving smaller mesoscale features such as isolated thunderstorms and heat island effects that the MMS network alone may fail to capture.

Horizontal Scale	Lifetime	Stull (1988)	Pielke (2002)	Orlanski (1975)	Thunis and Bornstein (1996)	Atmospheric Phenomena
10 000 km	1 month	Macro	Synoptic Regional	Macro- α	Macro- α	General circulation, long waves
2000 km	1 week			Macro- β	Macro- β	Synoptic cyclones
200 km	1 day			Meso- α	Macro- γ	Fronts, hurricanes, tropical storms, short cyclone waves, mesoscale convective complexes
20 km	1 h	Meso	Meso	Meso- β	Meso- β	Mesocyclones, mesohighs, supercells, squall lines, inertia-gravity waves, cloud clusters, low-level jets, thunderstorm groups, mountain waves, sea breezes
2 km	30 min			Meso- γ	Meso- γ	Thunderstorms, cumulonimbi, clear-air turbulence, heat island, macrobursts
200 m	1 min			Micro- α	Meso- δ	Cumulus, tornadoes, microbursts, hydraulic jumps
20 m	1 s	Micro	Micro	Micro- β	Micro- β	Plumes, wakes, waterspouts, dust devils
2 m	1 s			Micro- γ	Micro- γ	Turbulence, sound waves
		Micro- δ			Micro- δ	

Figure 2.3. Atmospheric scale definitions, adapted from Thunis & Bornstein (1996). The mean minimum distance between stations has been added to express how station density correlates with the various scales of atmospheric phenomena.

Figure 2.4 further illustrates the benefit of being able to access such a dense network. The number of professional Met Office stations across London is limited in comparison to the dozens of CWS sites. It is clear that were an urban heat island study conducted for London, with reliable measurements, the CWS could capture district level temperature variations that the Met Office network is too sparse to resolve. Another use of the CWS data may involve feeding it into the data assimilation scheme of a high resolution numerical forecast model (Section 2.2.2); the ~1.5 km square grid cells used in the Met Office's UKV configuration have been overlaid. By including CWS stations many more of these grid cells will actually contain an observation. Assuming the observation is reliable, then the atmospheric state of these cells would be better characterised. Even with the citizen stations included it is strikingly apparent that the

land observing networks are struggling to keep up with the pace of recent improvements in model resolution, such as those seen in the UKV.



Figure 2.4. Spatial distribution of weather station networks over the London conurbation on the 1st June 2014. The figure shows the professional Met Office MMS network amongst the citizen networks WOW and Weather Underground. The overlaid grid is the ~1.5 km grid used in Met Office’s short-range forecast model (UKV). UKV grid coordinates were first converted from an equatorial lat/lon projection to WGS84 using an internal Met Office function, before reprojected in esri’s ArcGIS software, along with the WGS84 formatted station locations, to the OSGB36 projection of the underlying LCM2007 land cover map (Morton, et al., 2011) shown.

2.1.2. Temporal resolution

The frequency at which citizen weather data is submitted is often as impressive as the spatial density. The frequency at which observations are uploaded range from once daily to, more commonly, intervals of 15, 10 or 5 minutes and even, in some cases, every minute (Figure 2.5). It is interesting that Weather Underground receives the majority of its data at 5 minute intervals whereas WOW users also commonly submit every 15 minutes. The reason for this difference is unclear as both websites allow data to be submitted at any of these intervals, but may stem from differences in the intermediary software used to submit a CWS’s observations to each website.

Note that for common variables such as air temperature, relative humidity, and wind speed, CWS along with their accompanying software submit point observations (as

supposed to averages). However, rainfall is generally submitted both as a daily accumulation, and as a rate per hour.

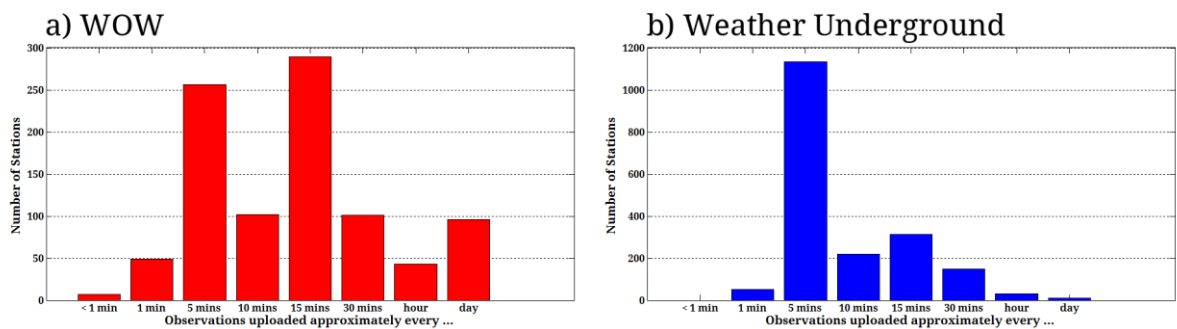


Figure 2.5. Number of UK stations uploading temperature observations to a) WOW and b) Weather Underground at the specified upload frequencies on the 1st June 2014. The observation frequency was derived after having web scraped a day's worth of observations from every station that uploaded at least 1 observation that day to wow.metoffice.gov.uk and www.wunderground.com.

The appeal of such frequent observations is their ability to capture short-lived weather phenomena, whether that be isolated heavy showers, sting jets, or fast moving squall lines. With the Met Office system now reporting every minute (Green, 2010) it is reassuring to see that citizens are capable of matching this. Credit is due to the web servers capable of receiving, storing and serving up these volumes of data. For example WOW had received 12.5 million observations just half a year after its launch in 2011 (Weather, 2012) and 150 million by the end of 2014.

2.1.3. Common automatic weather stations

When it comes to buying a new weather station, citizen observers have a wide range of station designs and manufacturers to choose from. With so many on offer they can choose one to suit their individual price range and needs. Figure 2.6 gives a sense of the weather stations on the market, and their popularity.

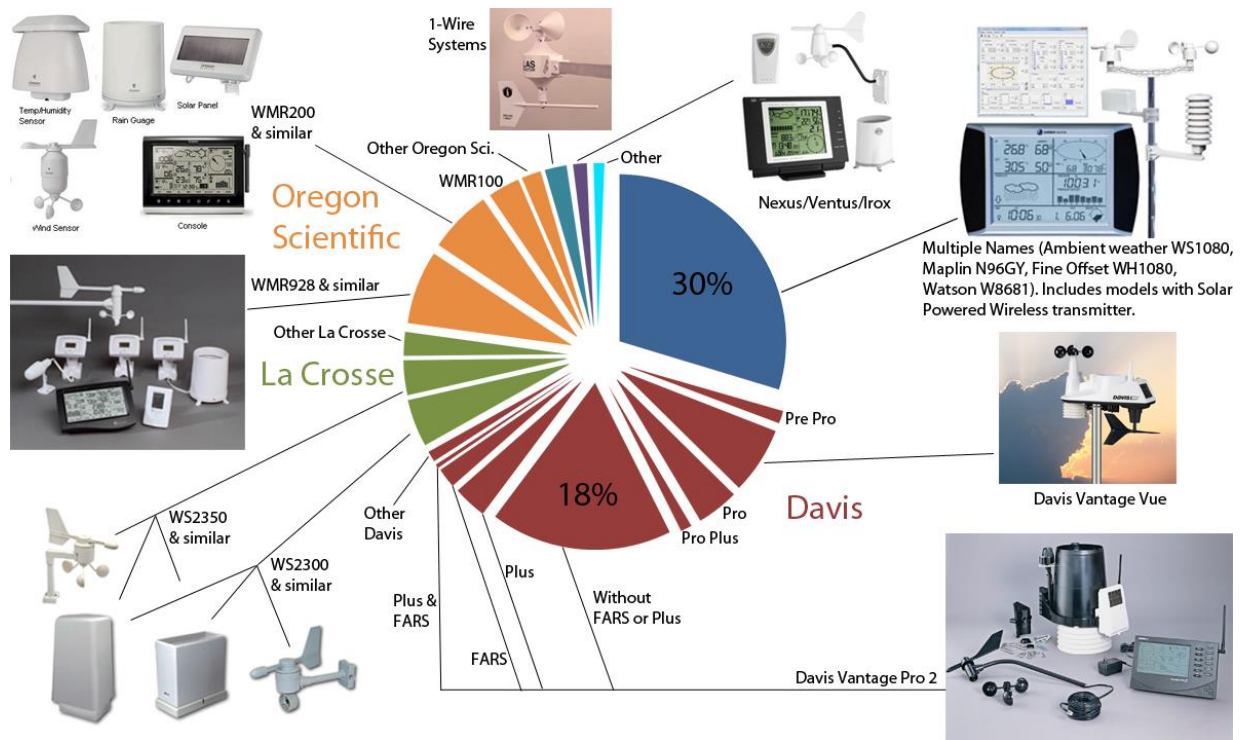


Figure 2.6. Weather station manufacturers and models used to automatically upload data to Weather Underground in February 2012. A total of 1353 stations were investigated, of which 16.6% were of unknown type and have been excluded from the diagram (Appendix 8.2). Davis ‘Plus’ models incorporate solar radiation and UV sensors, ‘FARS’ stands for Fan-Aspirated Radiation Shield.

A similar distribution of stations is evident on the Met Office’s WOW website (Appendix 8.3), with the Fine Offset WH1080 and Davis VP2 each accounting for approximately one third of total number of known station models. Common stations range in cost from £50 to in excess of £1000. The majority comprise an outdoor sensor suite with an indoor electronic console to display and log the observations. Once the console is connected to an internet-enabled computer or specially configured router, the data can be uploaded automatically using software such as Weather Display, Cumulus, Meteobridge or WeatherSnoop.

The variety of CWS station models in operation presents a challenge when it comes to bias correction. As Section 3.2 demonstrates, the different station designs can produce very different bias characteristics. As such, learning the station type, whether that be from metadata or the data itself, is crucial to constrain the bias correction we should apply.

2.1.4. CWS metadata

Metadata describing a CWS station is as important as the observations it makes. It can detail everything from the location of the site, the sensors being used, to any biases

the owner may have already spotted. For the end user to have access to this metadata the citizen must compile and share this information. Thankfully websites such as WOW and Weather Underground encourage citizens to fill out a metadata form when signing a CWS up to their site.

Location Attributes

Please describe the **location attributes** of the observation equipment at this site.

Exposure

Temperature

Rainfall

Wind

Urban Climate Zone

Reporting Hours

1

C

D

U

5

U

Figure 2.7 Screenshot from wow.metoffice.gov.uk showing the rating system a citizen is encouraged to complete when registering a station on WOW. A list of possible values for these attributes is shown in Appendix 8.5.

A feature of WOW is that citizens can add more than just basic metadata about their site. Most notably WOW employs a site grading scheme whereby citizens can rate the quality of their temperature, wind and rainfall measurements, along with their station's exposure and the Urban Climate Zone index, which when combined give the site an overall star rating. Figure 2.8 shows the distribution of these star and temperature ratings, indicating that WOW hosts observations of widely varying quality.

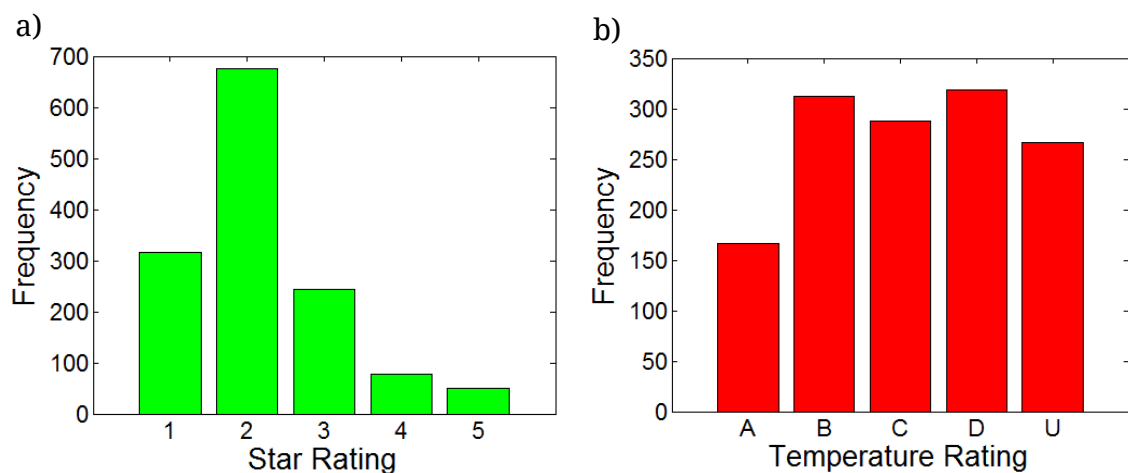


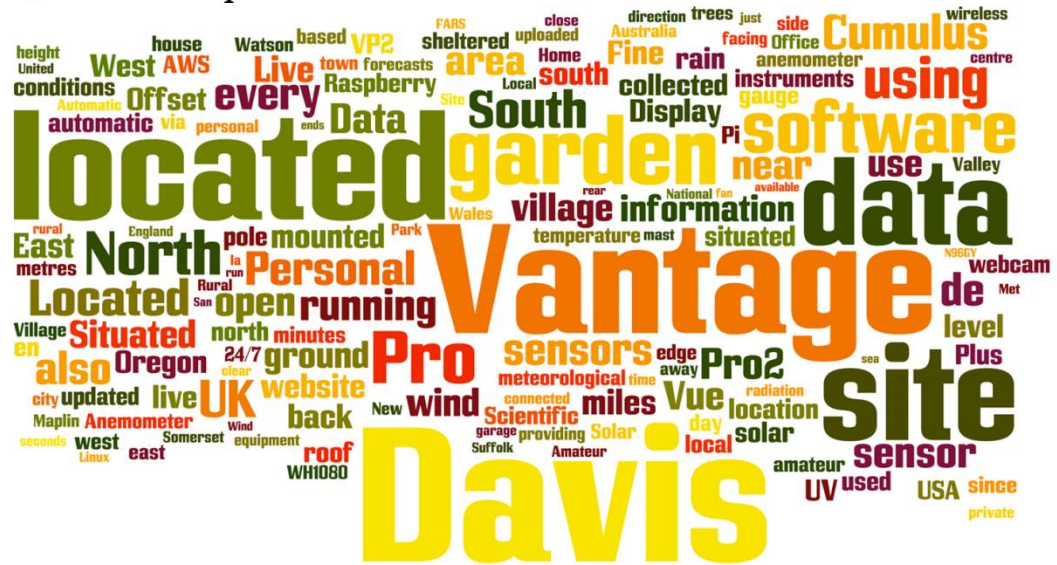
Figure 2.8. Distribution of a) Star and b) Temperature ratings on the Met Office's WOW website as rated by the owners of each station. Ratings of 5 and A respectively correspond to the highest standard. A full breakdown of these ratings is given in Appendix 8.5. To retrieve these ratings for every WOW users a web scraper was written to extract this metadata from the webpage of each station. Sample size = 1361 stations (includes non-UK stations).

The same grading system was first employed by the Climatological Observers Link (COL), with the Urban Climate Zones derived by Oke (2004). This type of metadata is very useful as it helps quantify the quality of the observations in a consistent manner.

For example a data user could use this metadata to quickly discard all stations whose temperature measurements fall below the standard required or to only select stations located in rural areas. In Section 5.2 we assess whether a station's rating corresponds to the degree of bias its data displays.

If the citizen wishes to add any other information they can do so in the two text boxes labelled *Site Description* and *Additional Information*. Figure 2.9 helps visualise the type of information commonly provided within these two text boxes. Words used more often appear larger in the figure. The figures show how these text boxes are frequently used to list what model of weather station is being used, something we rely on in Section 5.4.1. Note that it is unclear where a WOW user should write their model of station, with manufacturer and model names appearing frequently in both sections. This should be made clearer. WOW users may also upload photographs of their sites, providing an instant sense of the siting and sensors being used.

a) Site Description



b) Additional Information

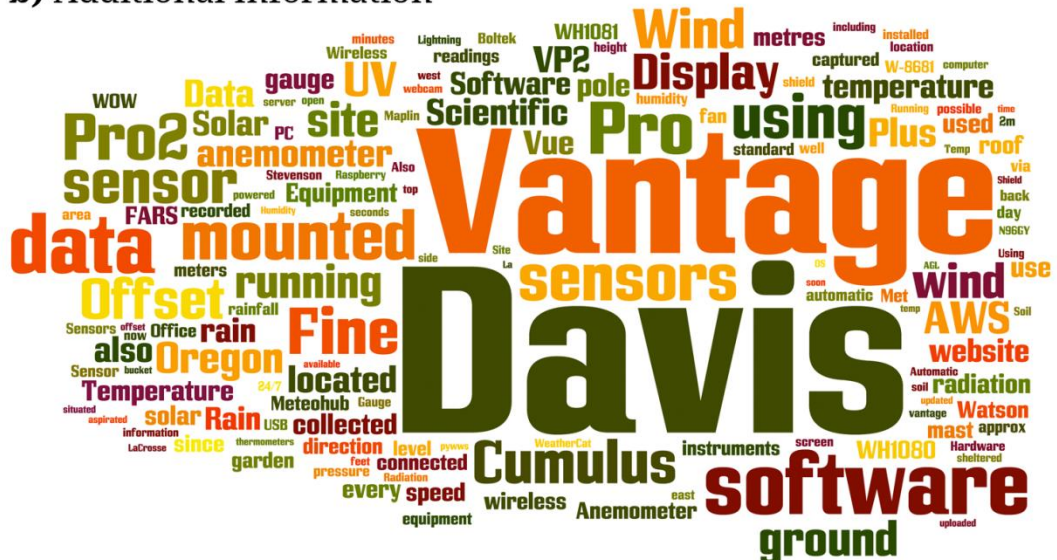


Figure 2.9. Word clouds visually representing the textual metadata provided by WOW members in the a) *Site Description* & b) *Additional Information* sections used to describe their stations to other members. Produced using the online tool at www.wordle.net. The size of the word indicates how frequently it is used by the WOW community as a whole. In November 2013, when this information was extracting from WOW using web scraping techniques (Section 2.1), only 640 out of 1100 WOW stations had any *Site Description* text available, and only 363 had listed *Additional Information*.

Weather Underground also has a small metadata section in which users can select their model of station from a drop down list, denote what surface type their stations sits above, along with the basic height, elevation and location information.

It is clear then that citizen observers have the option to compile a variety of metadata about their site; information that is crucial for users to interpret the data. However, as

detailed in Section 2.3.4 there are often issues with CWS metadata that make it difficult to assess the accuracy of a CWS site.

2.2. Applications of CWS data

2.2.1. Previous applications

CWS data has already been used in several applications. In each, data from multiple stations were collated and analysed for a given scientific application. For example, the Citizen Weather Observer Program (CWOP – www.wxqa.com) is an initiative which demonstrates how, with adequate quality controls, CWS data can be made available to external organisations and weather services. With over 7000 member stations in the USA, and over 10,000 worldwide, the majority of which are CWS, the worldwide popularity and value of such observations is clear. Here in the UK, studies by Muller (2013) and Illingworth (2014) show the willingness of citizens to participate in making observations. With the aid of social media and by engaging primary schools they collected snow depth and rainfall data respectively at high spatial resolutions. Muller (2013) received 170 snow depth observations over the Birmingham region, and Illingworth (2014) established a network of 6 primary schools in Birmingham, and 4 in Manchester.

Practical applications have also been seen in the Netherlands; where two separate studies - Wolters and Brandsma (2012) and Steeneveld, et al., (2011) - used CWS data to quantify the urban heat-island (UHI) effect. The first study selected suitable citizen stations based on strict criteria regarding their siting and exposure, accepting data from only 10% of the available stations. The second removed stations that showed unphysical behaviour in the timing of the maximum UHI, and used the uncertainties quoted by the station manufactures to guide their confidence in the data. In our approach we instead only discard observations that are clearly gross errors (Section 5.6.2 details what we define as a gross error), opting to correct rather than discard biased observations. The above approaches are also relatively labour intensive, while by contrast we demonstrate how the biases and uncertainties can actually be learnt automatically from the data itself. By demonstrating such an approach to correcting CWS data we hope that the appeal of the data for use in other applications will increase.

2.2.2. Potential applications

A promising application of CWS data is to better constrain the near-surface initial conditions in high-resolution numerical weather prediction models. One such model

is the Met Office's UKV configuration which represents reality as a series of 'grid cells' 1.5 km square. As was illustrated in Figure 2.4, the vast majority of these cells do not actually contain a weather station to initialise their near-surface state, but by incorporating the 1000+ citizen stations in the UK, more of these cells could be better characterised. In situ surface observations could, therefore, have an important role to play in high-resolution data assimilation. With a greater density of stations comes the opportunity to better forecast phenomena such as deep convection, which is highly sensitive to small-scale variations in surface temperature and moisture (Browning, et al., 2007). Although, as noted by Browning, et al., (2007), it would still prove difficult to accurately forecast individual storm cells, even with ~1 km model resolutions.

Citizen observations may also have a role in locally adjusting the output of model forecasts. Much post-processing involves correcting site-specific forecasts (Moseley, 2011) by 'learning' a systematic correction to the numerical forecast based on a historical sequence of observations. Unsurprisingly, but conveniently, the locations of citizen sites in the UK and Ireland follow a similar pattern to population density, meaning that corrections can be learnt particularly easily for areas where many people can benefit from the improvement.

2.2.3. Useful variables

Incorporating citizen observations into weather forecasting systems will be more beneficial for certain observed variables than for others. For example, citizen sensors that record temperature are relatively accurate, often to within 0.2 °C when well sited and ventilated (Huband, 1990). Observations of relative humidity are associated with greater uncertainty, but humidity can vary significantly in the distance between one professional station and the next (Alves & Biudes, 2013), so the use of citizen sites to fill in the gaps could be very beneficial.

Precipitation is also highly localised (Bohnenstengel, et al., 2011), and so supplementing our country's rain gauge network with citizen gauges could better capture this spatial variability, with particular benefits for applications such as flood prediction. Snow cover and snow depth are difficult variables to measure at unmanned sites, but observations made manually by citizens and uploaded to the internet could help to compensate for the dwindling number of staffed professional sites.

Observations of pressure are less crucial, since coherent structures in the surface pressure field typically occur on the larger mesoscale and synoptic scales, which are well resolved by the Met Office network. However, smaller-scale structures,

sometimes associated with high-impact severe weather, are occasionally observed (e.g. Browning and Hill (1984); Clark (2011); (2012)), and higher-density measurements of surface pressure would be useful in such cases.

Finally, wind direction and speed measurements taken by citizens can have many inherent problems, which make them less suitable for use at a regional or national level. Even for stations with separate wind sensors, the influence of the turbulent urban-boundary layer can present serious difficulties if the goal is to measure the mean wind direction and speed of the surrounding few kilometres (Oke, 2004).

2.2.4. A need to quantify uncertainty

The aforementioned uses of CWS data are only viable if the data collected by the CWS are actually representative of the ‘true’ value of each weather variable. With any citizen data there is a degree of uncertainty as to whether the data matches reality; quantifying this uncertainty and any associated biases is crucial to users wishing to incorporate the data into their application. As explained in Section 5.5, the ‘true’ values we wish to observe depend on the scale at which the given application needs to resolve.

As mentioned previously, one use of CWS data could be to feed it into data assimilation schemes with the aim of improving the initial conditions in a numerical forecast model. A key concept within data assimilation schemes is that they require uncertainty (error variance) values for each observation they ingest in order to properly weight this new incoming data against background weather forecasts; these values tend to be specified within the observation error covariance matrix (Bouttier & Courtier, 1999). These values are often specified according to knowledge of the instrumental characteristics of the given sensor, which are estimated using collocated observations. However only a few studies have performed collocated tests such as this on CWS sensors (Burt (2009); Burt (2013); Jenkins (2014)), which is one reason why we performed the field study detailed in Section 3.1. This observation error covariance matrix should also include the variance of representativeness errors as well. With CWS observations in particular, the associated observational uncertainty is likely to be non-stationary in time and in space. Therefore it is important to develop a model capable of quantifying these fluctuations.

Other applications that rely on associated uncertainty estimates are studies of climate change and urban heat islands. These applications aim to detect temperature changes, whether that be in space or time, of the order of only a few degrees Celsius. As uncorrected CWS biases can reach well in excess of these values (Section 3.2) it is

imperative that users have access to reliable uncertainty estimates to inform their use of the data.

2.3. Sources of uncertainty

Uncertainty about CWS' data can arise from any of the following five sources:

1. **Calibration issues** – a CWS sensor may not be perfectly calibrated. Perhaps it was biased before installation, or it has drifted over time.
2. **Design flaws** – often the design of a CWS makes it susceptible to inaccurate readings, particularly during certain weather conditions.
3. **Communication and software errors** – can produce gross errors as well as missing data.
4. **Metadata issues** – incomplete or inaccurate metadata make data interpretation difficult.
5. **Representativity error** – the representativity of an observation can be wrongly estimated. This can lead to an observation being inappropriately used to represent a spatial area or time window that is different from what it really sampled.

We look at each of these sources in greater detail below – providing examples for each. This list has been derived both from errors exhibited during the intercomparison field study (Chapter 3), and from publications such as Strangeways (2003), CWOP (2005), Overton (2007), Burt (2012) and WMO (2010) who describe such errors and issues in greater detail than that supplied here. They also offer suggestions on how to combat them. These 5 issues affect professional sites too, however with greater experience and resources professional organisations have the means to mitigate their impact. Conversely, although many citizens are aware of these issues, most do not have the means to abide by the recommended solutions.

2.3.1. Calibration issues

Calibration is crucial for all meteorological instruments. It ensures that not only are the sensors measuring the world around them accurately, but that their measurements are consistent and comparable with other calibrated sensors. Unfortunately the calibration of CWS sensors is frequently neglected – partly due to ignorance of its importance, but also because it can be difficult and/or expensive for citizens to perform. Take, for example, the capacitive sensors found in many automated weather stations used to measure relative humidity. Over time they can 'drift', especially when exposed to long periods of high humidity; with a tendency to

read higher (and thus ‘wetter’) by around 1–2% per year (Visscher & Kornet, 1994). This problem of drift is not isolated to citizen sensors; it occurs in professional systems as well (Ingleby, et al., 2013). The difference is that professional stations are regularly subject to recalibration using specialised equipment – an option rarely available to citizens. Even a simple zero-check for thermometers in iced water proves difficult, as many of the thermometers used in automated citizen weather stations are mounted on a circuit board, which can easily be damaged when submerged in water. In Section 5.6 we demonstrate an approach that compensates for these possible calibration biases by learning, over time, a correction to counteract their effect.

2.3.2. Design flaws

The design of popular CWS can be quite different from that of standard professional equipment. Although many manufacturers appear to imitate professional setups the final product will have differences, whether obvious or subtle, capable of producing relative biases. These differences arise as a manufacturer tries to save costs to target the citizen market, whilst also making their product novice-friendly and aesthetically pleasing. Some designs even indicate a lack of basic meteorological understanding.

A common design flaw evident in some of the most popular CWS is a poorly designed radiation shield prone to overheating. The overheating often results from a combination of poor radiation shielding and insufficient ventilation. The result is a warm bias under strong insolation, exacerbated during calm conditions. Williams, et al., (2011) noted a warm bias in citizen temperature data uploaded to Weather Underground when compared against nearby Met Office data, with overheating a likely cause. The Fine Offset WH1080 (Figure 1.1) is also available in the colour black, and thus provides an example of a design whose aesthetics compromised its accuracy, as the station is prone to an increased absorption of solar radiation and thus overheating. In Section 5.4.2 we propose that the various different designs of shielding fall into 7 common classes, each of which has its own bias characteristics.

Some stations, such as the Davis Vue (Figure 1.1), combine all their sensors into one integrated unit. This makes it easy for a novice to set up, but because the sensors cannot be separated it makes it virtually impossible to mount each individual sensor at the correct height and exposure. On one hand, the rain gauge should ideally be located 30 cm above ground level to avoid under-catch, which worsens with distance from the ground (Green, 1970), whilst on the other hand the anemometer should be at a standard height of 10 m, or even higher when surrounding terrain is not flat nor

unobstructed (WMO, 2010). Finding a suitable middle ground ultimately leads to some trade-off.

Many CWS rain gauges are also poorly designed with respect to professional gauges. They may be too small, with small tipping buckets, shallow sides, and without sharp edges around the rim, each of which can cause errors (Overton, 2007). A gauge with a larger horizontal surface area would catch more rain allowing its tipping buckets to increase in size (i.e. a greater volume of water is required to make them tip) whilst still keeping the resolution of the measurements the same. A resolution of 0.2 mm, or even 0.1 mm, is recommended (WMO, 2010). If a greater volume of rainfall is required to cause a tip then the gauge will be less sensitive to detritus that may build up in the buckets, or to error caused by the buckets not fully emptying as they tip.



Figure 2.10. The Fine Offset WH1080 rain gauge. Note its small rectangular shape (51 × 111 mm), and small tipping buckets.

The field study results, in Section 3.2, show evidence of bias resulting from many of the design flaws discussed here. Crucially this section then attempts to quantify and parameterise these effects.

2.3.3. Communication and software errors

Once a sensor has converted its observation into an electrical signal any error introduced to the data as it travels from the station, through the electronic console's memory, into the elected software package and up onto the selected data hubs falls into this category. These errors are characteristic of the automatic CWS we are focusing on and do not tend to befall manual weather observations. For example, many of the common CWS transmit wirelessly from the sensor suite to the electronic display console. This link may experience interference from nearby devices

submitting on the same frequency, potentially causing spikes in the data. If this communication link is not stable, perhaps because the station is sited on the limit of the wireless range, then the data may also contain a lot of missing values. Potential software errors may include a mismatch in the units used by the software uploading the data and that used on the online data hub receiving the information, or perhaps an incorrect timestamp may be allocated to each observation.

Many of the errors in this category are gross errors. A basic quality control step to pre-process the data, such as that used on WOW data in this project (Section 5.6.2), should remove many of these. The more subtle errors, however, remain in the data. The bias correction model (Section 5.6) must therefore be constructed to sensibly handle these types of error. It should also be set up to expect large amounts of missing data.

2.3.4. Metadata issues

As citizen sites do not have to follow the same strict regulations as their professional counterparts there can be a lot of variety in the exposure, siting and instruments in use. Thus accurate and detailed metadata is arguably of greater importance for CWS sites than it is for professional networks. As Section 2.1.4 described, citizen observers do have options when it comes to sharing metadata about their station. However, at present there are many issues with CWS metadata.

The first issue is that many of the websites that share CWS data only display a limited amount of metadata, preventing thorough citizens from sharing any extra metadata they may have collected. There is also no standardised metadata format on every website. For example, WOW and Weather Underground use very different metadata forms. Weather Underground for example has a drop down menu for citizens to select their station model from a list, which provides some consistency between users; whereas on WOW users must take the initiative to write their station model into one of the text boxes. However, only 70% of WOW users add information to these text boxes and of these only 60% clearly write what model of station they own. This is not the only example that demonstrates that many citizen observers are unaware of the importance of providing accompanying metadata. For example on WOW around 15% of citizens neglected to update any of their site ratings from the default values, making it very difficult to assess the exposure, siting and instruments at those sites. In Section 5.5.1 we detail an approach to combat this problem – capable of estimating the exposure and Urban Climate Zone remotely using only the coordinates of the site to begin with.

When it comes to specifying the elevation of their site, almost a quarter of WOW users neglected this metadata field. Of those that supplied information, it is clear that some users have provided the wrong elevation as evident in Figure 2.11. It appears that several users have even submitted their elevation in the wrong units. There are also many sites for which the user denoted an elevation of 0m, but for which the DEM would suggest the height is much larger. This elevation problem is dealt with in Section 5.6.2.

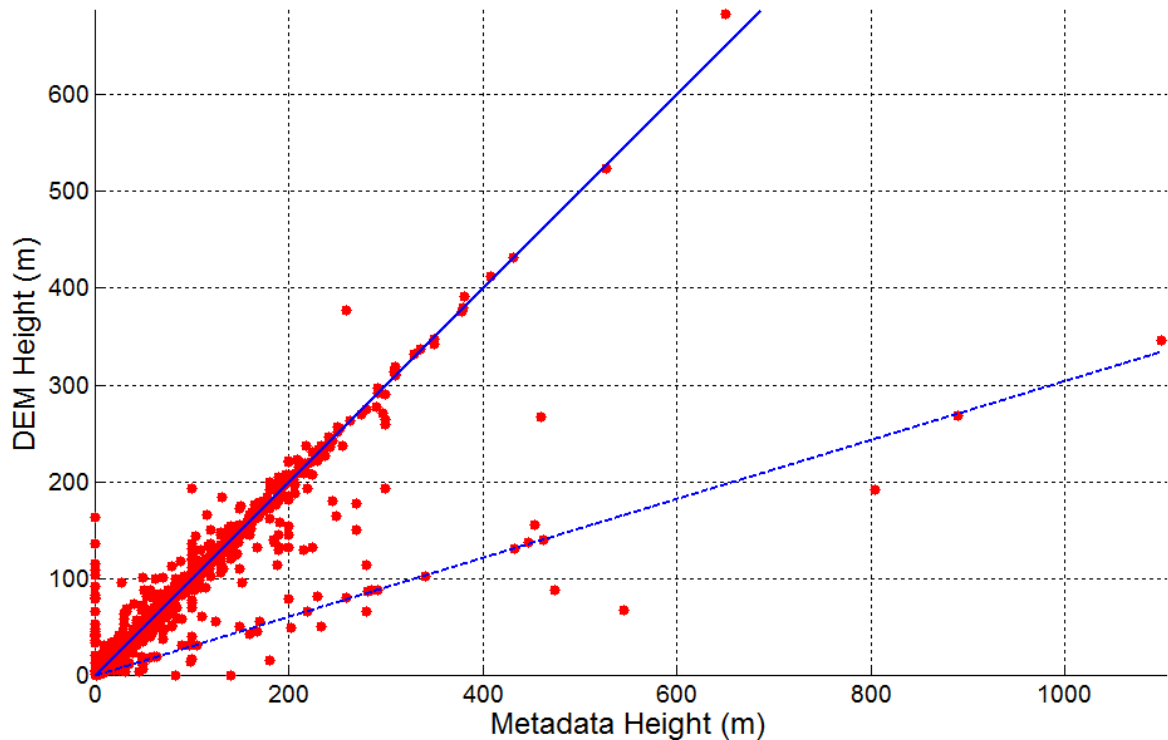


Figure 2.11. Comparison of user-contributed station elevation against the height for that location extracted from the GMTED2010 digital elevation model (DEM; Section 4.4.2). Solid line indicates the 1:1 line. The dashed line shows the 1:1 line were the metadata height given in feet.

Thankfully both WOW and Weather Underground do at least enforce the provision of station coordinates. They also both give the option of uploading an image of the site; however, few users upload such pictures, and interpretation of such images cannot be automated.

Ideally a standardised metadata form should be included on all data hubs that share CWS data; based upon best practises such as those outlined in Muller, et al., (2013). This would make it much easier for users of the data to assess whether a given citizen station is suitable for their application.

2.3.5. Representativity error

Representativity continues to be a challenging concept in meteorology, which makes quantifying whether a CWS observation is characteristic of a given areal extent a difficult task. The representativeness of an observation is the degree to which it accurately describes the value of the variable needed for a specific purpose (WMO, 2010). Representativity errors arise when the CWS observations sample real-world phenomena at spatial scales that are different from those we wish to resolve in a given application. Using temperature observations as an example – a CWS mounted over grass in highly vegetated and sheltered garden may not give a fair representation of the temperature over a surrounding neighbourhood which is otherwise highly developed, with a complex urban morphology primarily covered by artificial surfaces such as paving, tarmac and brick (surfaces with a higher heat capacity). Therefore unless a given application wants to resolve scales equivalent to that individual garden then the observations from such a station may not be representative.

Representativity is largely a function of siting and exposure. A well sited CWS resides in an area similar to its surroundings. However, in heterogeneous landscapes, finding such a suitable spot is difficult, even more so for a citizen observer who wishes to mount their station in their own garden. Many gardens also have a poor exposure, therefore rather than capturing the larger scale weather the station is instead heavily influenced by localised microclimatic effects, which may even be specific to that individual garden. Jenkins (2015) even showed that in a domestic garden temperatures can vary by several degrees Celsius depending on the location of the thermometer within the garden. Sheltered sites pose other threats too. Nearby trees and buildings can alter the catch of rain gauges simply by presenting an obstruction or by altering wind speeds (Guo, et al., (2001); Strangeways, (2007)). Sheltered sites with low wind speeds may prevent sufficient ventilation through the small passively-aspirated screens used on many CWS to house the thermistors and capacitive humidity sensors. This can exaggerate any design flaws that already restrict ventilation leading to inaccurate and time-lagged observations (Harrison, 2011). Warm or reflective surfaces nearby, such as houses, may also induce biased temperature readings (Oke, 2004).

A few previous projects have attempted to quantify the representativity of professional stations. An ‘inverse footprint’ approach was employed by Orłowsky & Seneviratne (2014) to define the spatial representativeness of European stations contributing to the ECA&D project. The larger the area surrounding a station in which neighbouring stations exhibited a strong correlation in their temperature anomalies,

the more spatially representative the station was thought to be. Such an approach could be applied to the stations used in our study; however the coastal geography of the British Isles would limit the ‘footprint area’ around coastal stations potentially causing an artificial drop in the quantified representativity. It is also important to consider the very local influences that can induce representativity error. For example, subtle changes in a site’s exposure (Brandsma, 2004), layout (Hewitt & Clark, 2013) or the presence of a nearby road (Kumamoto, et al., 2012) can all influence a thermometer’s readings. High-resolution numerical models have also been used to quantify representativity errors (Waller, et al., 2013), showing that the horizontal errors are correlated and more significant for specific humidity than temperature.

Each thermometer placed above a surface ‘sees’ only a portion of its surroundings. This source area, or ‘footprint’, that the station senses is incredibly difficult to define. It depends on the height of sensor as well as the characteristics of the turbulent motion that transports the sensed air to the thermometer. As Figure 2.12 demonstrates the source area is not symmetrically distributed around the sensor location. Due to wind effects it is elliptical in shape, and aligned in the upwind direction. For screen-level temperature measurements it is thought to have a typical radius of 0.5 km although factors such as sensor height, surface roughness, building density and atmospheric stability can all alter the size and shape of this footprint through time (WMO, 2010).



Figure 2.12. A conceptual representation of source areas (footprints). Image from WMO (2010), Chapter 11. The dark shaded ellipses show the theoretical source area for sensors responding to the turbulent transport such as thermometers, 50% or 90% of the signal comes from the area inside the respective ellipses. They are dynamic, moving with wind speed and direction, and atmospheric stability.

Many CWS cannot follow the strict exposure guidelines imposed on their professional counterparts. Because of this the area for which they are representative tends to be much smaller and less suitable for larger scale applications. Also many CWS are located in urban areas prone to heterogeneous land cover types and complex building configurations. The result is source areas distorted from the theoretical ellipse and that comprise of a mix of different land cover types, each with different thermal properties.

Representativity errors have the potential to exceed instrumental errors; indeed even in the absence of any instrumental errors a station can still contain representativity errors if used in an application resolving a different scale. It is therefore crucial we attempt to quantify for which scales we believe a station is representative. Therefore, in Section 5.5 we explore a possible link between a CWS's exposure and Urban Climate Zone and these errors, and in Section 5.6 we detail how representativity is handled within our bias correction model.

2.4. Summary

Organisations and individuals looking to CWS data as a potential source of low cost information to feed into their application should be impressed with the volume of data available, and with its spatial and temporal density. They should also be conscious that bias can enter the data from a variety of different sources. Sensible users must realise that attempts should be made to quantify this bias and associated uncertainty so that they can use the data with confidence.

3. Parameterising station bias

This chapter focuses on identifying, quantifying and parameterising instrumental bias inherent to CWS data. Without any evidence of these instrumental biases it proves difficult to differentiate these artificial biases from natural spatial variations when looking at real CWS data. This chapter explains how an intercomparison field study was used to collect *a priori* knowledge of the bias and uncertainty characteristics of different models of CWS. In Chapter 5 we apply what has been learnt here to model the bias in the real CWS data from WOW.

3.1. Field study design

3.1.1. The test site

A year-long intercomparison field study was performed using seven CWS collocated alongside professional Met Office equipment at the University of Birmingham's 'Winterbourne No. 2' site (Figure 3.1). The site located in Edgbaston, Birmingham, is part of the MMS network submitting minute-resolution data.

The Met Office instruments installed at the site include:

Air Temperature: A standard Met Office 100 Ω platinum resistance thermometer (PRT) mounted within a passively ventilated Stevenson screen ~1.2m above a grass. The resistance is sampled every 15 s from which a 1 min mean temperature is calculated. The measurement accuracy of calibrated PRTs is equal to ± 0.05 °C (Clark, et al., 2014). PRTs are usually replaced every 8 years. PRTs offer greater accuracy along with a more stable calibration in comparison with thermistors (Burt, 2012); all seven CWS we tested use thermistors.

Relative humidity: A Rotronic HygroClip MP100H, also mounted within the Stevenson screen. The Met Office quote typical uncertainties of $\pm 2\%$ per year deployed. HygroClips typically drift by 1% per year (Ingleby, et al., 2013). Like the PRTs, the HydroClip's report 1 min averages (made up of four 15 s samples). Recalibrated annually. Explained further in Section 3.2.3. Rotronic (www.rotronic.co.uk) quote a response time of 10 seconds (to perform 63% of a humidity change).

Rainfall accumulation: A Munro R100 series 0.2 mm tipping-bucket rain gauge. Tips are recorded at 1 minute resolution. Gauge rim ~30 cm above grass.

Global radiation: A Kipp and Zonen CMP11 pyranometer measuring global radiation. Mounted approximately 1.5 m above the ground. Records a 1 min average from 1 s samples, with a typical hourly uncertainty of < 3% (www.kippzonen.com).

Henceforth these reference Met Office instruments are referred to simply as *MMS*. The *MMS*'s instruments are calibrated on a regular managed cycle (approx. once a year) and abide by World Meteorological Organisation (WMO) standards (WMO, 2010); we assume therefore that they can be used as a well characterised reference against which the seven CWS can be verified. However, although this site acts as a suitable reference, the *MMS*'s stations are not immune to problems. For example, passively ventilated Stevenson screens suffer from increased uncertainty at low wind speeds (Harrison, 2010), while tipping-bucket rain gauges, such as the Munro R100, can display biases when verified against standard manual rain gauges (Burt, 2012). Fortunately, over the yearlong study period the Munro R100 we used read just +1.1% higher than a Met Office MK II 'five-inch' manual rain gauge and +1.6% higher than another, newer, Munro tipping-bucket gauge, both collocated at the site. With this relatively small bias and a virtually complete annual dataset we can use its readings with reasonable confidence. The site is somewhat sheltered, and therefore unsuitable for Met Office wind measurements, but fortunately the University of Birmingham maintains a set of instruments at the site, including a 7 m mast with an anemometer and wind vane manufactured by Vector Instruments. The site's sheltered nature is similar to that of many CWS. Further site metadata is provided in Appendix 8.6.

The study took place from 1 September 2012 to 31 August 2013. Having envisaged that the type and magnitude of the CWS' bias would depend on synoptic conditions, which vary through the year, a full year's field study was undertaken.

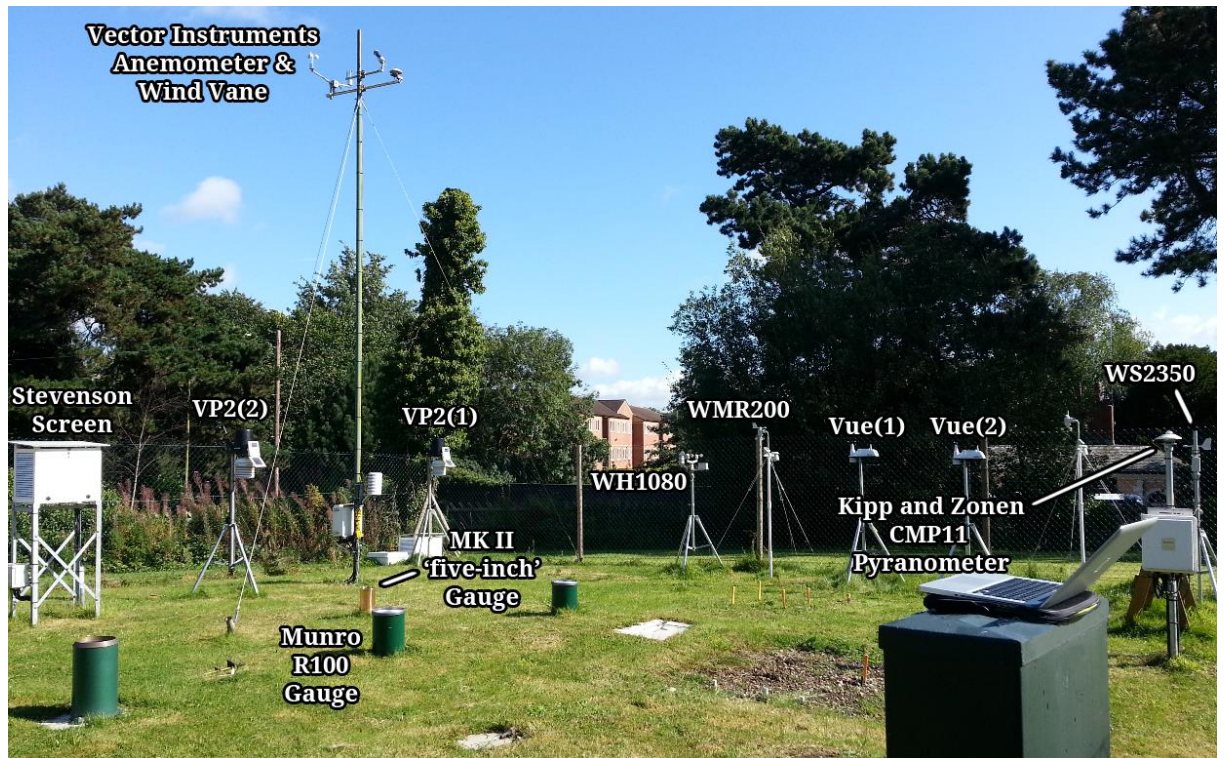


Figure 3.1. The Met Office's Winterbourne No. 2 weather station. The site includes sensors operated by the Met Office, the University of Birmingham, and the seven CWS being tested as part of this study. Photograph taken facing in a North-Easterly direction.

3.1.2. Tested citizen weather stations

The seven CWS comprised five different models of weather station, chosen because they are among the most popular automatic stations used by citizen observers (Bell, et al., (2013); Appendix 8.3). Details of the stations are summarised in Table 1, with images of the sensor suites shown in Figure 1.1. For Davis Instruments' Vantage Pro2 (VP2), Vantage Vue, and Oregon Scientific's WMR200 two of the same station were installed; with the aim of identifying biases and errors common to a particular model. However, the second WMR200 was decommissioned in early November 2012 when its wireless transmission began to interfere with that of the first station. We are confident that there was negligible interference before this point. Only a single Fine Offset WH1080 and La Crosse WS2350 were deployed because of fears of similar interference. With hindsight, the La Crosse instruments could have used wired communications and Jenkins (2014) used two Fine Offset devices simultaneously without issue. Like most CWS, every station comprised an outdoor sensor suite and an indoor electronic console to display and store the data. Observations were downloaded from the console to a laptop on a weekly basis. All CWS' and MMS's temperature and humidity sensors were mounted approximately 1.5m above grass. Although the rims of the MMS's rain gauges were roughly 30 cm above grass, the heights of the CWS' gauges were set as recommended in their manuals, ranging

between 1 m for the WS2350 and WMR200 to 1.5–2 m for the other CWS. This was done with the aim of replicating the height most citizen observers would mount their gauges at, assuming of course that the citizens follow the manual guidelines. The positions of the CWS and MMS sensors at the site were not changed during the entire yearlong study.

Table 1. Summary of the 7 CWS tested as part of this field study.

Station Nickname	Station Manufacturer	Station Model	Price ^a (Approximate)	Software used to download observations	Temporal resolution (minutes)	Transmission Frequency ^e (seconds)	Time until memory full at this temporal resolution (days)	Rainfall increment (mm)
VP2(1)	Davis Instruments (www.davisnet.com)	Vantage Pro2 FARS ^b	£890	WeatherLink (www.weatherlink.com)	10	10, 50, 20	18	0.2
VP2(2) ^c	Davis Instruments	Vantage Pro2 FARS	£890	WeatherLink	10	10, 50, 20	18	0.2
Vue(1)	Davis Instruments	Vantage Vue	£390	WeatherLink	10	10, 50, 20	18	0.2
Vue(2)	Davis Instruments	Vantage Vue	£390	WeatherLink	10	10, 50, 20	18	0.2
WMR200	Oregon Scientific (store.oregonscientific.com)	WMR200	£350	Virtual Weather Station (www.ambientweather.com)	10	60	291	1.016
WS2350	La Crosse (www.lacrossetechnology.com)	WS2350	£100	Heavy Weather (www.heavyweather.info)	60	8	7	0.518
WH1080	Fine Offset ^d (www.foshk.com)	WH1080	£70	EasyWeather (www.foshk.com)	10	48	30	0.3

^a Prices include accompanying software, but not mounting accessories such as tripods. Only the WMR200 comes with a mounting pole as standard. Prices include VAT. Source: www.weathershop.com and www.maplin.co.uk (as of Jan 2014).

^b FARS stands for Fan Aspirated Radiation Shield.

^c The VP2(2) had been in the field for approx. 1 year before installation at Winterbourne No. 2. All other stations were brand new.

^d Fine Offset manufacture this station but it is frequently sold under many different brand names including Maplin, Watson, and Ambient Weather.

^e The manuals of the stations tested do not directly specify the internal sampling procedure used, i.e. whether point samples or averages are recorded. Their manuals do however specify the transmission frequency, i.e. the rate at which the electronic console is updated with observations from the outdoor sensor suite. We saw no evidence of any further averaging of temperature, relative humidity, or rainfall accumulations once the observations arrived at each console before being saved to its internal memory. For Davis stations individual variables have different transmission frequencies, shown are the frequencies for temperature, relative humidity and rainfall accumulation respectively.

Table 2. Key statistics from the field study

Statistic Station	Air Temperature (°C)			Relative Humidity (%)			Dew Point (°C)	MSLP ^a (hPa)	Rainfall	Total No. of observations and percentage missed
	Mean Bias (Day and Night)	Mean Bias (Day time ^b)	Mean Bias (Night time)	Mean Bias ^c (all conditions)	Mean Bias (Wet conditions, >90%) ^d	Mean Bias (Dry conditions, <=90%) ^d	Mean Bias	Mean Bias	Absolute and percentage difference from the MMS yearly total of 842.4mm	
VP2(1)	+0.2 (0.2)	+0.1 (0.2)	+0.3 (0.2)	+2.7 (2.9)	-1.3 (1.2)	+3.6 (2.3)	+0.7 (0.6)	+1.7 (0.3)	-83.4 mm (-9.9%)	52558 (< 0.1%)
VP2(2)	+0.2 (0.3)	+0.1 (0.3)	+0.3 (0.3)	+0.4 (3.1)	-2.2 (3.1)	+1.0 (2.7)	+0.3 (0.5)	+1.2 (0.3)	+94.8 mm (+11.3%)	52558 (< 0.1%)
Vue(1)	+0.1 (0.3)	+0.2 (0.3)	+0.1 (0.2)	+2.7 (2.1)	-0.2 (0.9)	+3.4 (1.7)	+0.8 (0.7)	+1.7 (0.6)	-22.6 mm (-2.7%)	52560 (0%)
Vue(2)	-0.1 (0.3)	+0.0 (0.3)	-0.2 (0.2)	+3.9 (2.0)	+1.1 (0.9)	+4.5 (1.6)	+0.8 (0.7)	+2.9 (0.8)	-28.6 mm (-3.4%)	52560 (0%)
WMR200	+0.8 (1.3)	+1.5 (1.4)	+0.1 (0.4)	-11.0 (6.3)	-2.8 (4.1)	-12.8 (5.2)	-1.7 (1.4)	+2.6 (1.5)	-43.8 mm (-5.2%)	48318 (8.1%)
WS2350	+0.9 (2.3)	+2.1 (2.5)	-0.5 (0.5)	-1.4 (5.2)	-1.3 (2.0)	-1.4 (5.7)	+0.9 (1.4)	+1.9 (1.1)	-100.0 mm (-11.9%)	8679 (0.9%)
WH1080	+0.5 (0.9)	+0.9 (1.0)	+0.0 (0.3)	+7.5 (3.2)	+5.1 (1.9)	+8.0 (3.1)	+2.3 (1.3)	+0.0 (0.7)	-203.4 mm (-24.1%)	52471 (0.2%)

Over the period 1 Sept 2012 through 31 August 2013, except for Relative Humidity and Dew Point whose statistics represent the period 16 May 2013 – 31 August 2013. The standard deviation of the difference is shown in brackets next to the values of mean bias.

^a Winterbourne No. 2 site does not have MMS MSLP readings, instead observations from the Coleshill MMS site 16 km away were used. CWS pressure readings were set to match the Coleshill reading at the start of the period, except for the WMR200 for which the MSLP correction is based upon the elevation the user enters into the electronic console.

^b Here the definition of daylight is when the MMS pyranometer (1 min resolution) reads greater than 0 W m⁻² at the time of the CWS temperature observation. Therefore night time is when the reading is less than or equal to 0 W m⁻².

^c CWS humidity observations have not been corrected for temperature biases, which may compound humidity biases.

^d As measured by the MMS humidity sensor.

3.2. Investigation

Here we discuss and interpret the observations collected, dealing with each weather variable in turn, namely: air temperature, humidity and dew-point temperature, and rainfall. Because pressure variations are well captured by the MMS network, and CWS' wind measurements are generally too localised for most applications, they are not examined in detail. None of the CWS tested measure solar radiation: only 'Plus' versions of the VP2s measure solar radiation. Table 2 summarises key statistics from the year-long field study.

3.2.1. Temperature

When the air-temperature measurements from the seven CWS were verified against the MMS's measurements there were significant biases (Table 2), with clear diurnal and seasonal patterns (Figure 3.2). The pattern is dictated by the hours of daylight, with changes in the magnitude, and sometimes the sign, of the bias between day and night.

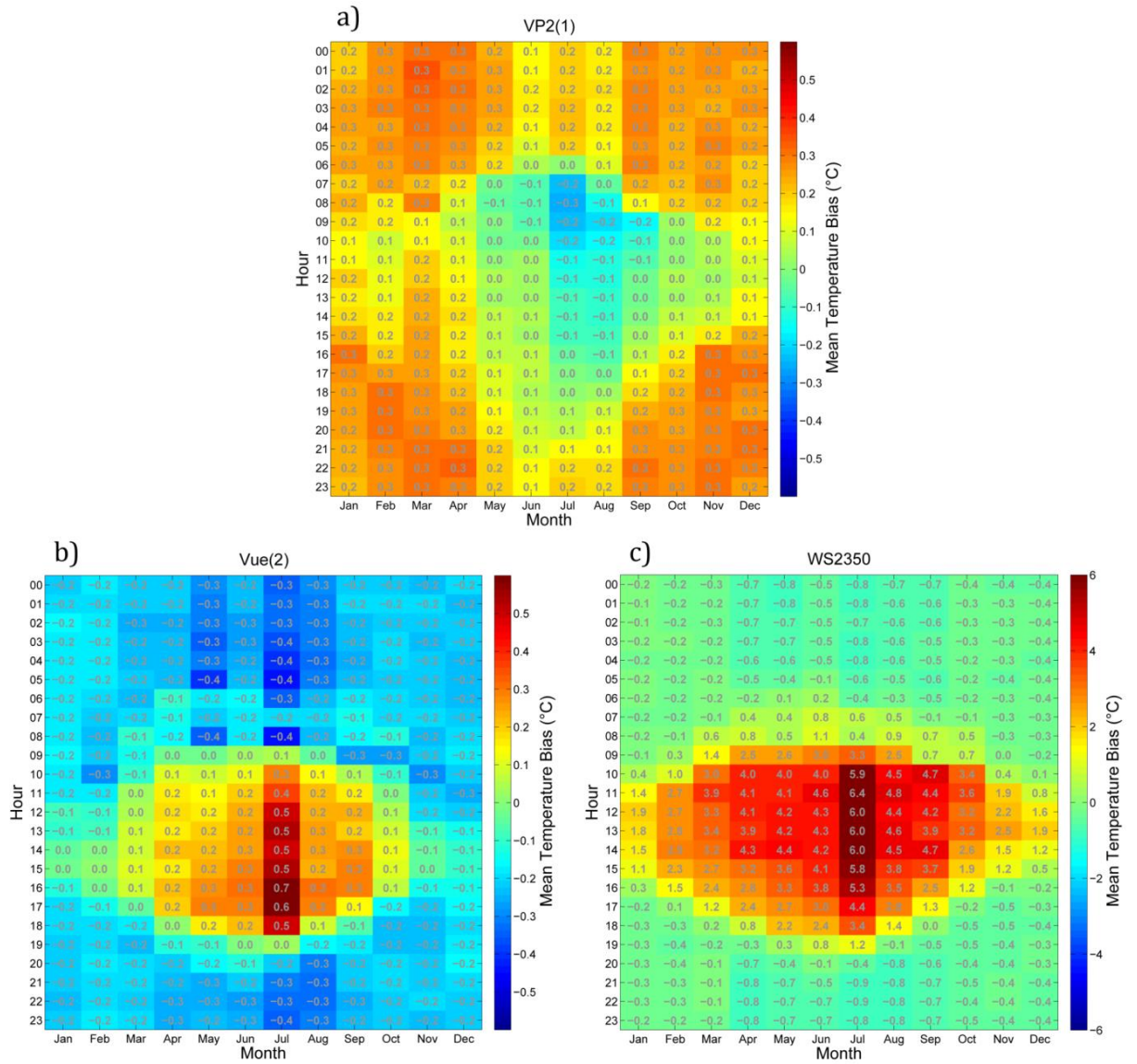


Figure 3.2. Mean temperature bias at different hours of the day (UTC) and months of the year for three of the CWS tested. (a) Davis VP2(1), (b) Davis Vue(2) and (c) La Crosse WS2350. Note the change in the colour scale for the final plot. The values written in grey are the mean bias of each cell. These values are simply the average of all bias values that fall within a given hour and month division. The VP2 and Vue sample every 10 minutes, whereas the WS2350 samples every hour.

The Davis VP2 and Vue stations show the closest agreement with the MMS's PRT, all with a relatively small mean bias and standard deviation. The two fan-aspirated VP2s show very similar results, both tending to read too warm, with overall average biases of +0.16 °C and +0.18 °C. Both show an increase in this warm bias at night and a decrease during the day. The two Vue stations, however, do not agree: Vue(1) tends to show a warm bias that is exacerbated in the afternoon, whereas Vue(2) consistently shows a cool bias, around -0.2 °C at night, which only changes to a warm bias between late morning and late afternoon for the warmer months of the year, as shown in Figure 3.2.

The temperature bias of the other stations is more significant, each showing a warm bias that dramatically increases during the day, and leads to a positively skewed distribution of bias for these stations. The pattern of the warm bias shown in Figure 3.2 for the La Crosse WS2350 station is typical for all these three stations, with a warm bias that peaks just after midday. During summer months the warm bias is even more pronounced, occurring for more hours of the day owing to the extended hours of insolation. These warm biases are well over 1 °C and can climb over 4 °C for the WMR200 and WS2350. To put these biases into context, the summer average daytime urban heat-island measured in London rarely exceeds 1 °C (Wilby, et al., 2011); as derived from the discrepancy between daily maximum air temperatures observations at St James's Park (urban) and Wisley (rural). Without accurate bias correction for CWS it would be almost impossible to accurately quantify such a daytime urban heat-island effect using some of the station types. At night the performance of these three stations is much improved, for example the WMR200 and WH1080 both display a small mean bias (standard deviation) of 0.05(0.4) and 0.03(0.3) °C respectively. At night urban heat-island effects are generally more pronounced as well, i.e. with larger urban-rural contrasts in temperature (Wilby, et al., 2011).

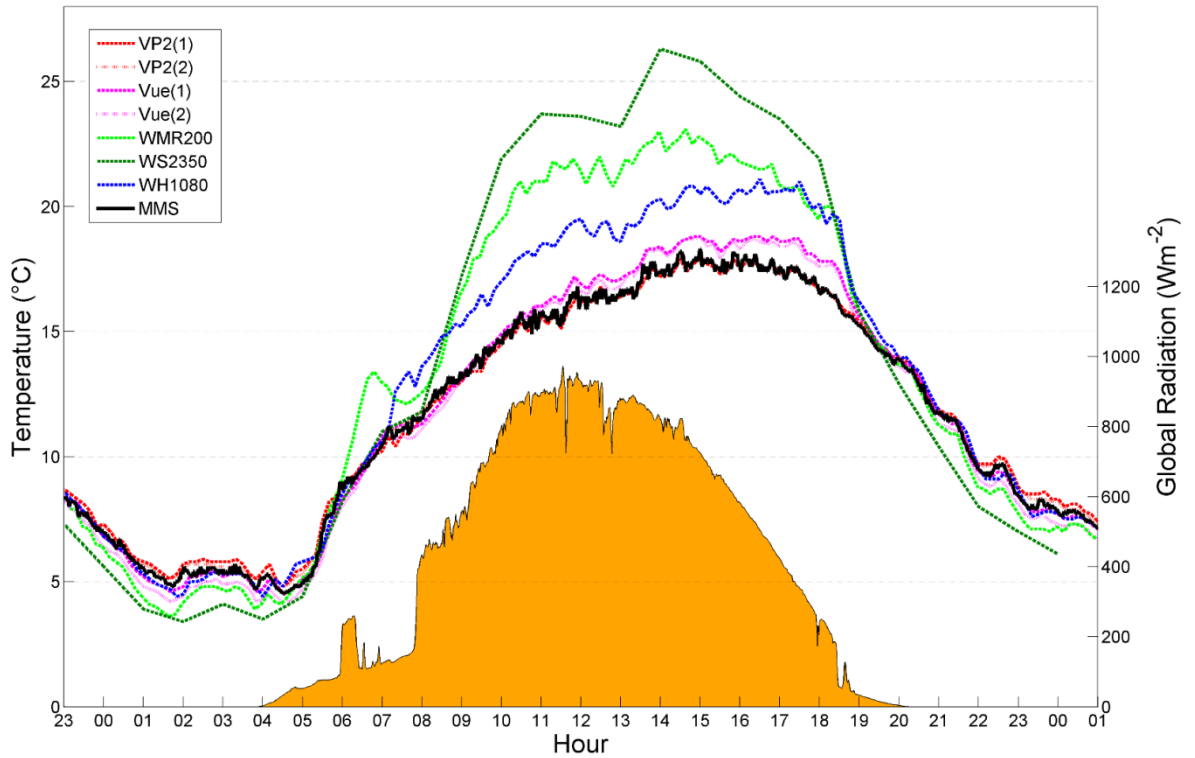


Figure 3.3. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 26 May 2013. A time series of MMS global radiation is shown in orange.

Figure 3.3 shows a temperature time-series plot for a typical day that highlights many of the patterns. For example, note the large daytime warm bias exhibited by the WS2350 station (and to lesser extents by the WMR200 and WH1080) and the warm bias of both Vue stations later in the day. The large step changes in global radiation during the morning highlight shading effects caused by the somewhat sheltered nature of the Winterbourne No. 2 test site. This highlights the impact of micro-scale siting. With nearby obstructions causing complex shading patterns that alter the strength of the radiation reaching the CWS, and thus the magnitude of the subsequent temperature bias. Such shading effects are common within domestic gardens, with the impact on low-cost thermometers shown by Jenkins (2015). The implications of these complex shading patterns on this study, resulting from each CWS being positioned slightly differently within the site, is unclear.

In Figure 3.4 we see that these dramatic warm biases can even occur in February when global radiation levels are much lower. Around 15:00 the WH1080 displays a dip in temperature not apparent in the other stations. This is perhaps indicative of a sheltering effect, i.e. the WH1080 is in shade whereas the WMR200 and WS2350 are not, and thus its warm bias is reduced. Accounting for such sheltering effects when bias correcting operational CWS observations is almost impossible.

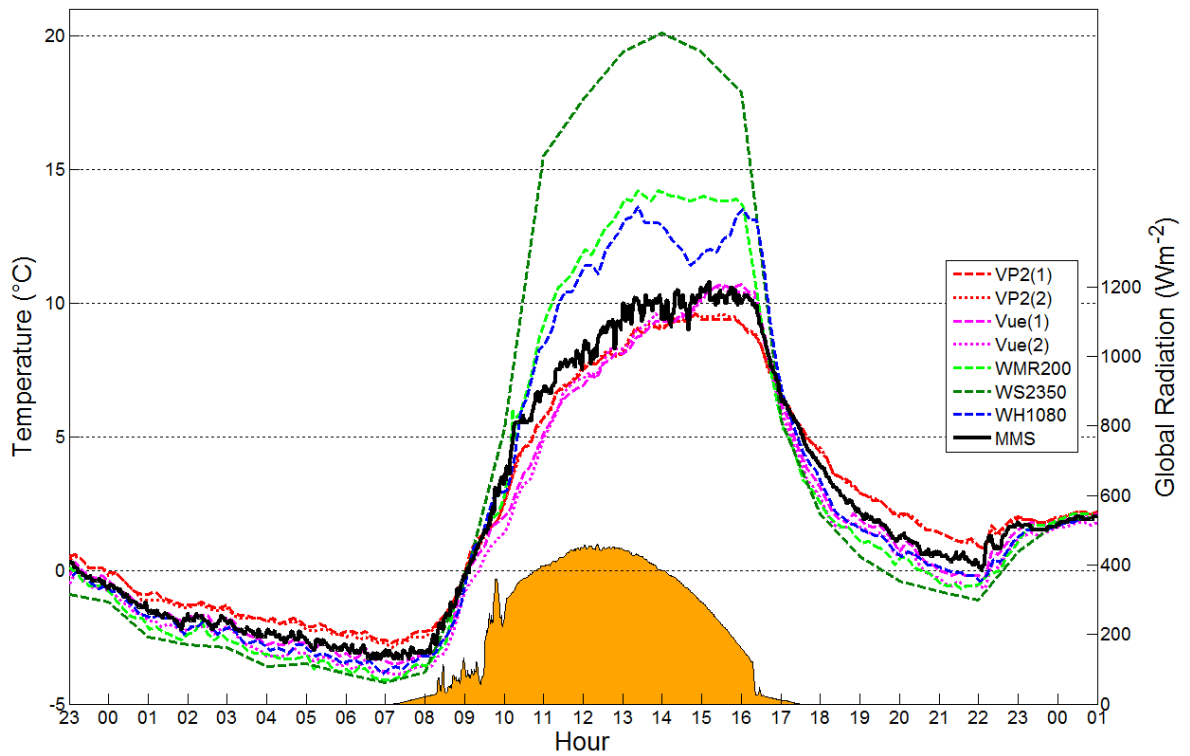


Figure 3.4. Time series plot of air temperature recorded by the seven CWS and the professional platinum resistance thermometer housed within a Stevenson screen for 19th Feb 2013. A time series of MMS global radiation is shown in orange.

The most obvious pattern to the temperature bias is the difference between day and night; a result of changes in the radiative balance. For the WMR200, WS2350 and WH1080 stations the main driver of their daytime warm bias is the strength of incoming solar radiation. Figure 3.5 shows how these three CWS exhibit a greater warm bias with increasing levels of global (direct + diffuse) radiation.

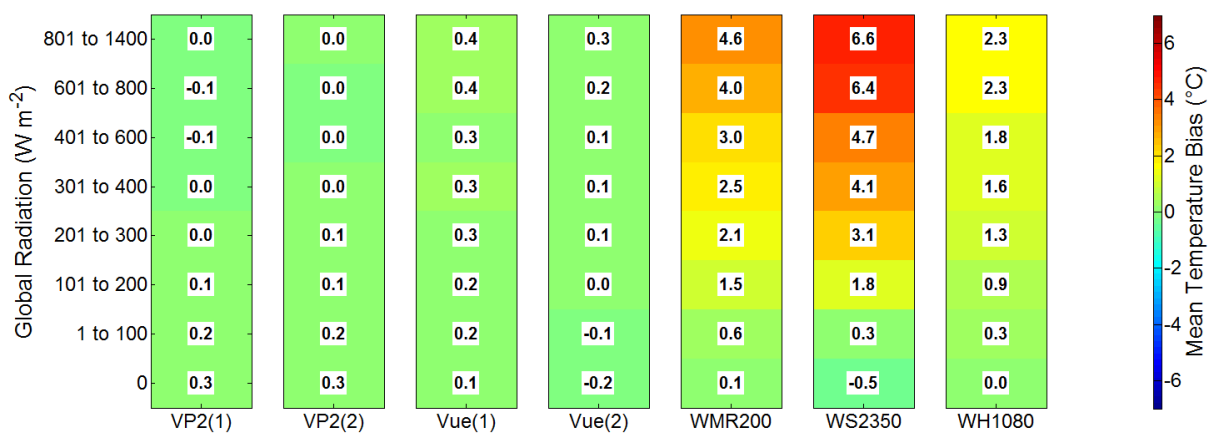


Figure 3.5. Temperature bias as a function of global radiation levels for the seven CWS tested. Global radiation observations less than 0 W m^{-2} were rounded up to 0 W m^{-2} .

For most stations this relationship between global radiation and temperature bias is well modelled by a 1st order linear regression model (Figure 3.6). However, for some

of the more bias-prone stations, particularly the WS2350, there is marginal improvement when a 2nd order polynomial approximation is used. This is because it appears that the bias begins to plateau as global radiation reaches its maximum. These linear and quadratic models were deemed suitable for what is a relatively simple relationship, and were favoured over LOWESS to avoid over-fitting to inherent noise.

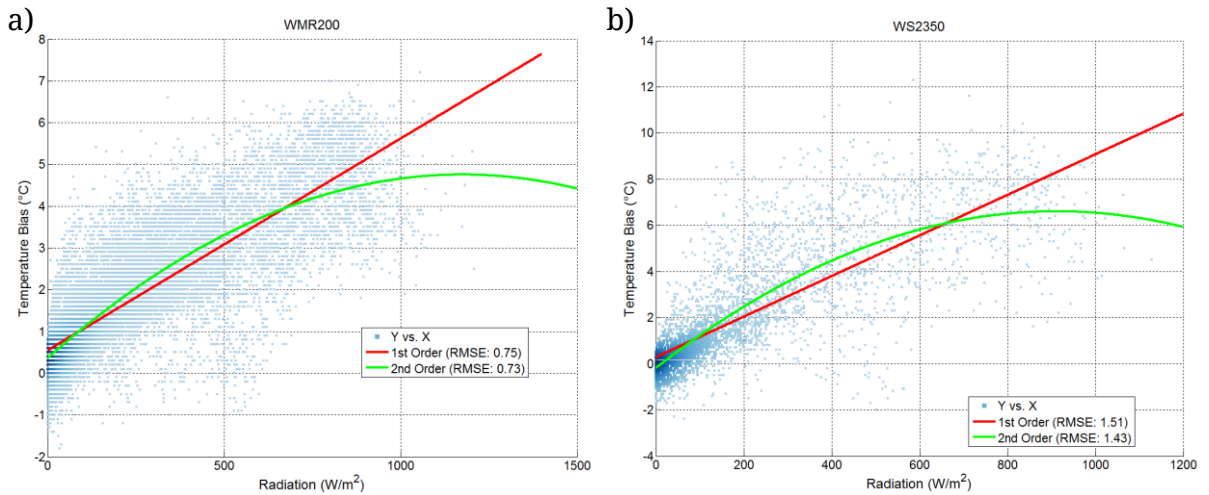


Figure 3.6. Relationship between global radiation and temperature bias for the a) Oregon Scientific WMR200 and b) La Crosse WS2350. The red and green lines show 1st and 2nd order regression models fitted to the data.

Wind speed also appears to influence this relationship. Note that in Figure 3.7 the warm temperature bias of the WMR200 station during high solar radiation conditions is exacerbated when the wind speed is low. The focus here is on wind speed, not wind direction, as we assume wind speed has a much greater impact on radiation shield ventilation.

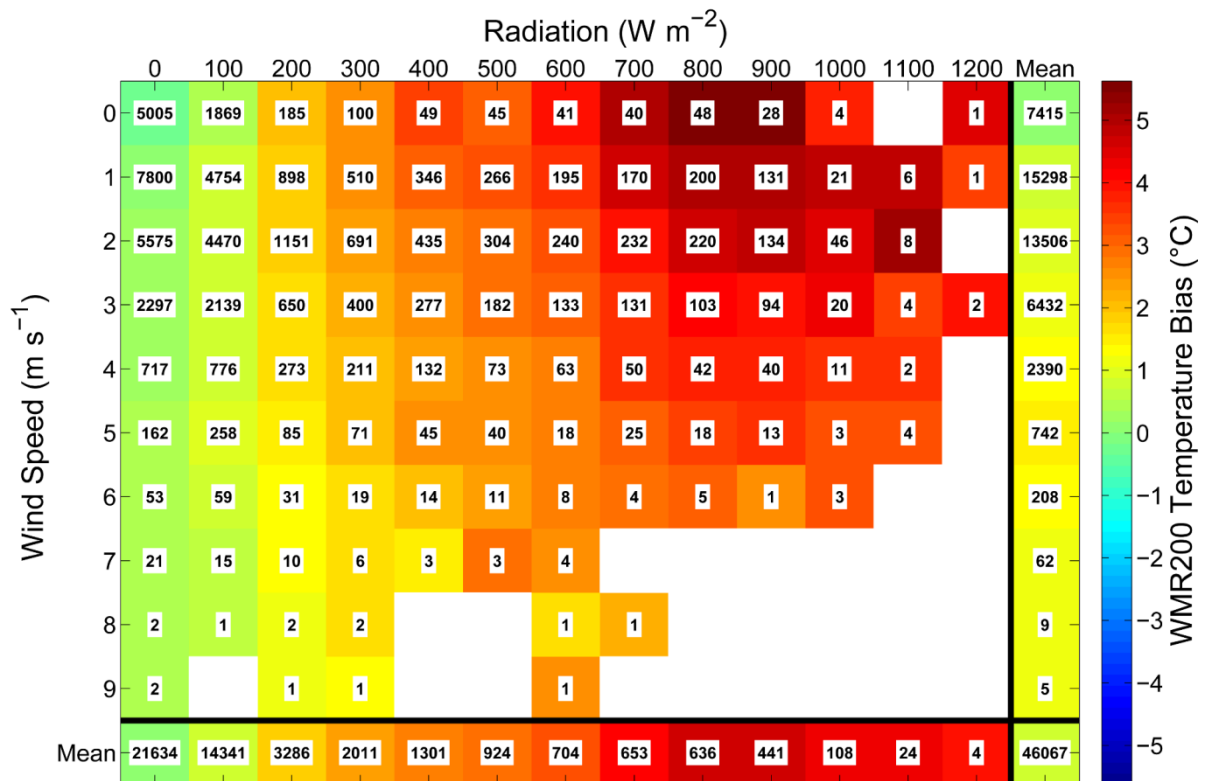


Figure 3.7. The WMR200 station's temperature bias as a function of wind speed and global radiation. The mean bias for a given radiation (wind speed) bin for all wind speeds (radiation levels) is shown along the bottom (right side). Here the number within each cell signifies the sample size.

All of the stations tested have some form of shielding to guard their thermistor from direct sunlight. Such shielding should also be ventilated to allow surrounding air to circulate through. It is apparent that some shields are more effective than others, as show in other studies (Hubbard, et al., (2001), Cheung, et al., (2010), Wheeler, et al., (2003)). Thermal images taken with a Flir i5 camera (www.flir.co.uk) under sunny conditions were used to highlight differences in the performance of the CWS shields (Figure 3.8). The images show that shielding of the WMR200 and WS2350 stations, which exhibit the largest biases under increased global radiation, display high infrared temperatures. This illustrates that their thermistor shielding is prone to overheating under sunny conditions, which heats the air inside the thermistor housing, thereby increasing the sensed temperature. This overheating is also a function of under-ventilation: the design of the shields of the WMR200 and WS2350 stations (Figure 1.1) makes sufficient ventilation difficult, so that the air within the shield warms, rather than being refreshed with ambient air from outside the shield. The upturned-plate design of the WH1080 station allows for better ventilation and is noticeably cooler than the WMR200 and WS2350 stations. However, the bias still displays a relationship with global radiation levels, perhaps due to its small size and

off-white colour. Jenkins (2014) also identified a relationship with solar radiation for the two WH1080 stations in their study.

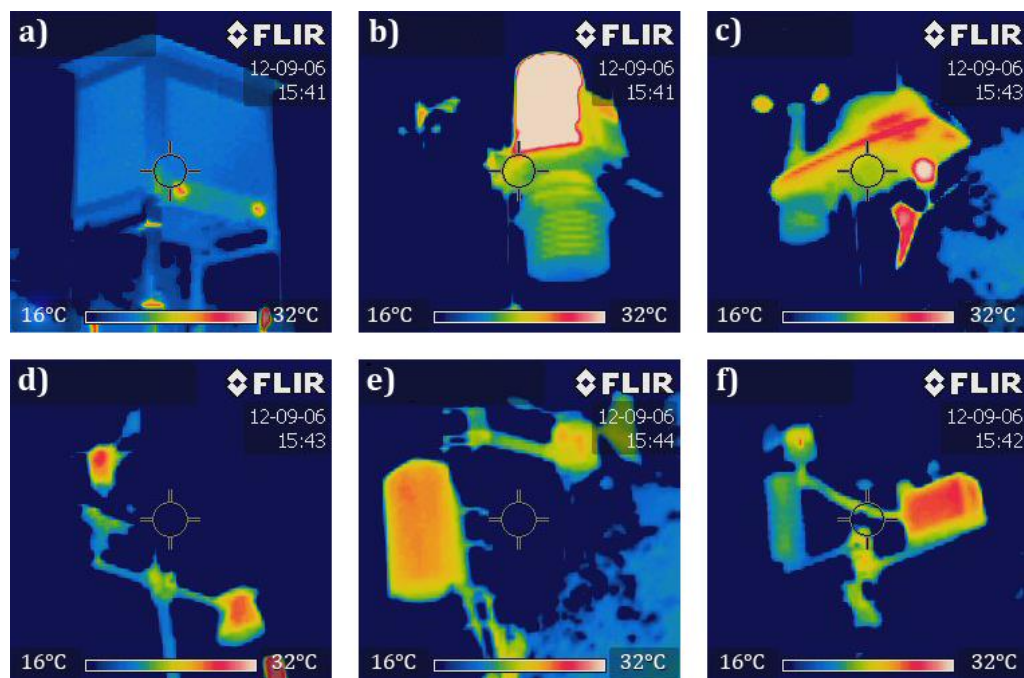


Figure 3.8. Thermal images taken from the southwest by a Flir i5 thermal imaging camera on a sunny summer afternoon: (a) Stevenson screen and (b) VP2, (c) Vue, (d) WMR200, (e) WS2350 and (f) WH1080 stations. All stations were in direct sunlight, and had been for several hours. The colour-scale is consistent. The white (hot) part of the VP2 station evident in panel (b) is its black rain gauge. For help identifying the parts of each station, cross-reference with Figure 1.1, but be aware of the change in perspective.

The two Davis models, the VP2 and Vue, display the coolest infrared temperatures in the thermal images, and their relationship with radiation is somewhat different. The VP2 stations were the only model we tested that included a fan-aspirated radiation shield (FARS). The fan is solar powered, and it is evident in Figure 3.2 that the VP2 station has the lowest bias during the day when this fan is active. Under sunny and calm conditions the aspirated VP2 stations probably provide a better estimate of the air temperature than the passively aspirated Stevenson screens, which are prone to increased uncertainty at low wind speeds (Harrison, 2010). During the night the fan is inactive, as this particular model of VP2 has no battery to power the fan when solar energy falls. As such ventilation can only occur passively, leading to a warm bias. The altered shield design, which incorporates active ventilation, has compromised the effectiveness of the passive ventilation at night. It is reassuring to see the similar performance of the two VP2 stations, providing confidence that the parameterisation would be similar for all stations of this type. However, having only tested 2 VP2s, there is no guarantee that this assumption is valid.

Davis's Vue stations appear to have a stronger relationship with outgoing longwave radiation than shortwave radiation. As the Vue station's radiation shield is mounted underneath its main body it is well shielded from incoming solar radiation, although its effectiveness may be compromised when the solar angle is low. It is when the land surface has warmed and outgoing radiation peaks, around mid-afternoon, that the station shows the greatest warm bias (Figure 3.2). Unfortunately, longwave radiation is not commonly measured at MMS's stations, so a proxy variable may have to be used to parameterise the Vue station's temperature bias.

3.2.2. Parameterising temperature biases

When it comes to bias-correcting these CWS temperature observations, Figure 3.9 demonstrates that with a reliable estimate of the global solar radiation level, as is available at the test site, it is relatively simple to apply an effective correction. Section 5.3 details how such a correction is made at locations where a collocated radiation sensor is absent. Here we used a simple multiple linear regression model in which radiation, wind speed, an interaction term and a constant term were used as predictors of the temperature bias, with radiation providing the most predictive power. When the model's temperature bias prediction was used to correct the observations there was a reduction in the mean bias and residual variance for all the CWS. Sometimes this improvement was marginal, for example, for the Davis stations, but for stations that exhibited large radiation biases, such as the WS2350, the improvement was large (as seen in Figure 3.9).

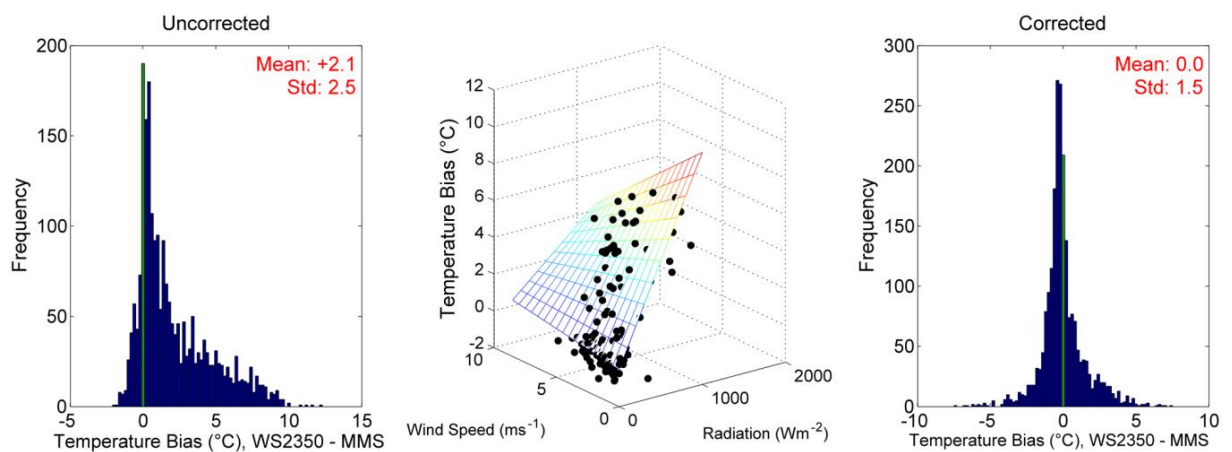


Figure 3.9. Demonstration of correcting CWS' temperature bias using a multiple linear regression model. The figure shows a histogram of the WS2350 station's temperature bias before and after the correction, along with a scatter plot of a sample of its observations overlaid with a grid of the learnt model. The data was randomly split in half to form the training and test datasets using daytime data only, for the WS2350 this resulted in 2222 training points and 2221 test points.

So far we have only considered the relationship between simultaneous CWS temperature bias and global radiation measurements. This assumes that the impact of changes in radiation on the temperature bias is instantaneous. When we consider that the impact on the temperature bias may lag behind changes in radiation, the strength of the relationship often improves. To incorporate this lagged effect one must include radiation observations that occurred before the time of the temperature bias, which raises the question of how many observations to use and how to weight them. The end goal here simple: to select and weight preceding radiation observations in a way that provides the best predictor of radiation-induced temperature bias. The better the predictor the more accurate the resulting bias correction is.

Figure 3.10 illustrates the different kernels that were tested to weight preceding radiation data (at 1 minute resolution) over a 60 minute window. The strength of the relationship varied little between window lengths of 30 through to 120 minutes, the key is to consider at least some previous observations. Weighting the observations in this way also helps to smooth the data, averaging out potential occasional noisy point observations.

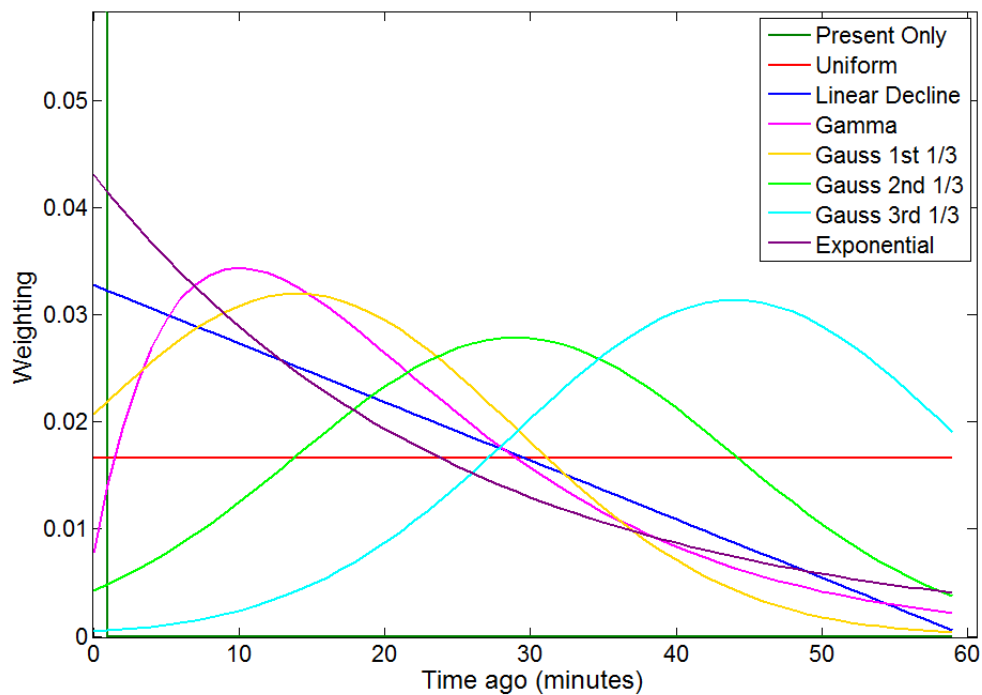


Figure 3.10. Kernels used to weight preceding global radiation (at 1 minute resolution) measurements over a 60 minute window.

Figure 3.11 and Figure 3.12 show the relative benefit of applying these different weightings. For all but the VP2s the relationship is significantly stronger when one of these weightings is applied relative to using present observations only, although the difference between the weightings is often very marginal. With these weightings

applied the correlation coefficients and R^2 values fall very close to 1 for the WMR200, WS2350 and WH1080, implying a strong positive correlation modelled well by the 2nd order regression model. As previously shown (Figure 3.6) a 1st order regression model would also perform well here; the 2nd order was used purely to make a fair comparison with the l_{Rad} column. l_{Rad} , explained in more detail in Section 5.3.3, stands for the logged radiation values that are used to improve the spatial interpolation of radiation observations to operational CWS locations. Before they are logged these observations are first weighted using the exponential kernel shown here (Figure 3.10); chosen because it has high R^2 values (Figure 3.12) and has the most physical meaning. What is reassuring is that even after the log transformation has been applied the relationship with the temperature bias is still strong. Thus when used in a 2nd order regression model to predict the temperature bias at operational CWS, radiation should provide an accurate estimate of the temperature bias, particularly for those poorer stations that display the largest biases. The assumption here is that we know what model of station is being used.

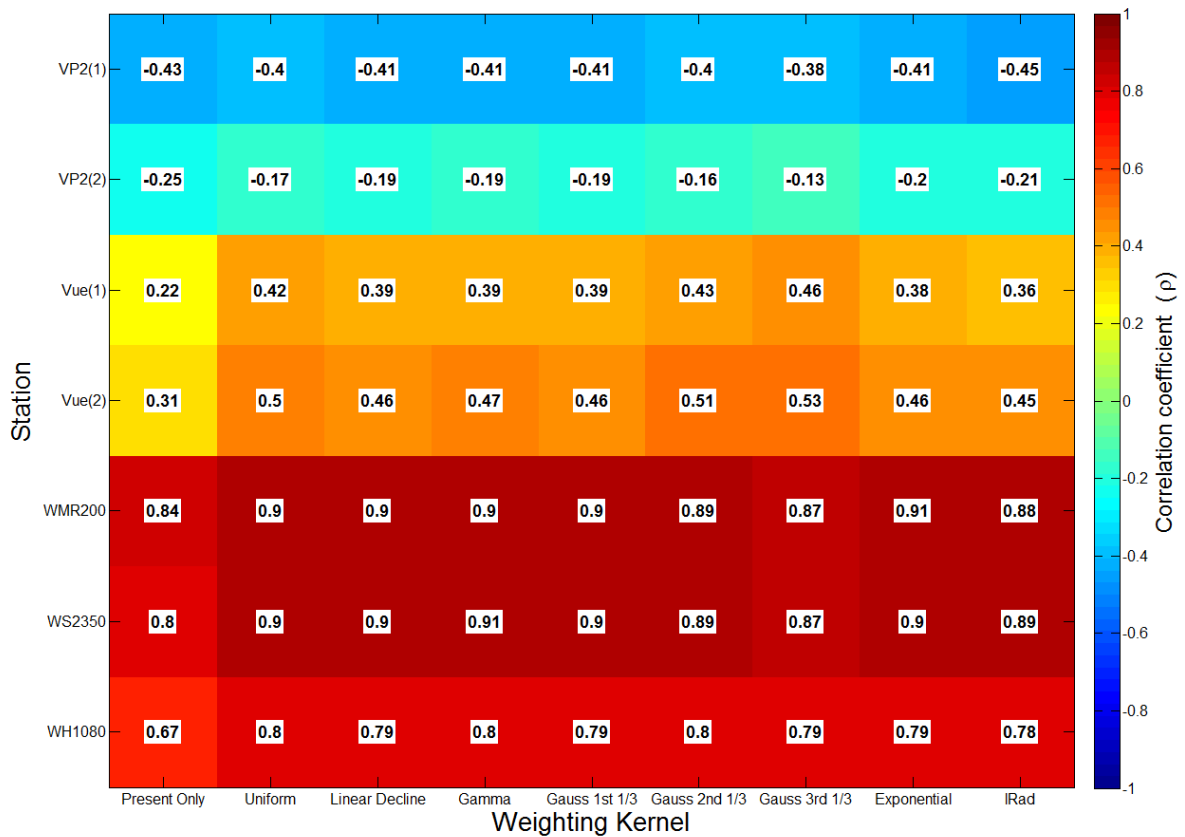


Figure 3.11. Correlation (Pearson's linear correlation coefficient) between each station's temperature bias and 60 minutes' worth of preceding global radiation observations (using 1 minute resolution radiation data) that have been weighted using a selection of different weighting kernels. Only observations during the day are used.

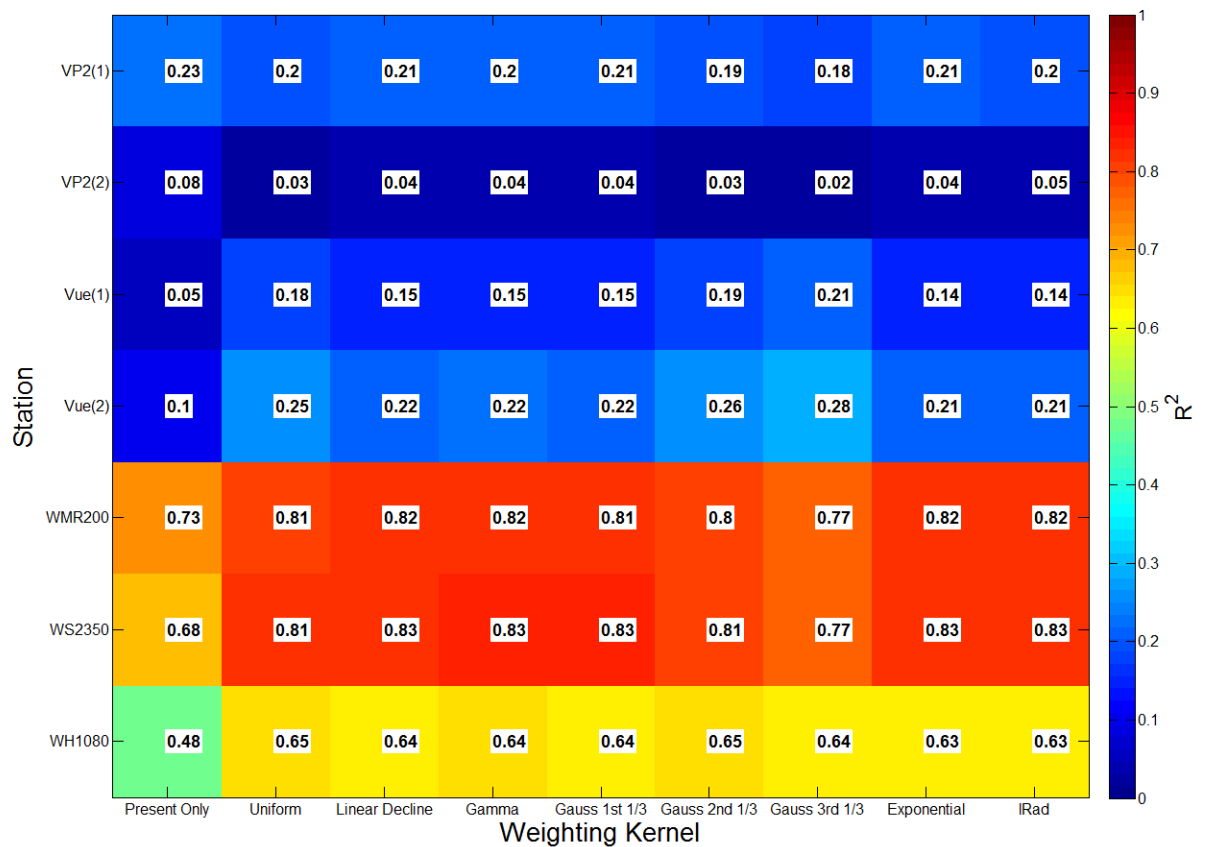


Figure 3.12. Relationship between temperature bias and 60 minutes worth of preceding global radiation observations (using 1 minute resolution radiation data) that have been weighted using a selection of different weighting kernels. The relationship is quantified using the R^2 statistic from a 2nd order regression model used to predict the temperature bias from weighted global radiation observations. Only observations during the day are used.

3.2.3. Relative humidity and dew point

All seven CWS tested exhibit significant relative-humidity biases when compared against the MMS's humidity sensor. The magnitude of these biases exceeds the typical range of discrepancy between different brands of professional hygrometers (Lacombe, et al., 2011), i.e. $\pm 3\%$ from the reference. It is important to note that there is some uncertainty associated with the MMS's humidity sensor, the Rotronics HygroClip. Figure 3.13 shows a time series of the Vue(1) humidity observations minus the MMS's observations. The sudden step change in the bias range in mid-May is because the Rotronics HygroClip was swapped for another as part of the site's calibration process. The Rotronics HygroClip that ran over the first 8.5 months tended to read much wetter than the CWS' sensors during conditions of high humidity, remaining at 100% for several hours if not days (Figure 3.14). The CWS' observations would rarely read as high as 100%. This explains the large negative biases shown in Figure 3.13 over this first period. The second Rotronics HygroClip showed no such tendency, exhibiting a much better agreement with a Vaisala humidity sensor run by the University of Birmingham at the site. In a separate field study, Ingleby, et al.,

(2013) found that Rotronics HygroClip sensors tend to drift by +1% to +2% per year at Met Office sites (although there is a lot of variability) and can be slow to recover from periods stuck at saturation. They estimate that an uncertainty of 2–3% for an operational Rotronics HygroClip is achievable under best conditions. As the first Rotronics HygroClip was deemed to have drifted wet, all following statistics and figures for relative humidity and dew-point temperature were produced using just the second Rotronics HygroClip as the reference sensor. Before their deployment, the HygroClips are calibrated in the Met Office quality assurance laboratory. Using an environmentally controlled chamber they quote a target calibration accuracy of $\pm 2\%$ over a range of 25–100% (Mander, 2012). Given that the second HygroClip was only in the field for, at most, 3.5 months its errors should fall close to this range. Were this field study to be repeated we'd recommend calibrating the HygroClip at the start of the year, and also after 6 months to correct for any drift, and to use multiple HygroClips at once to better quantify measurement uncertainty.

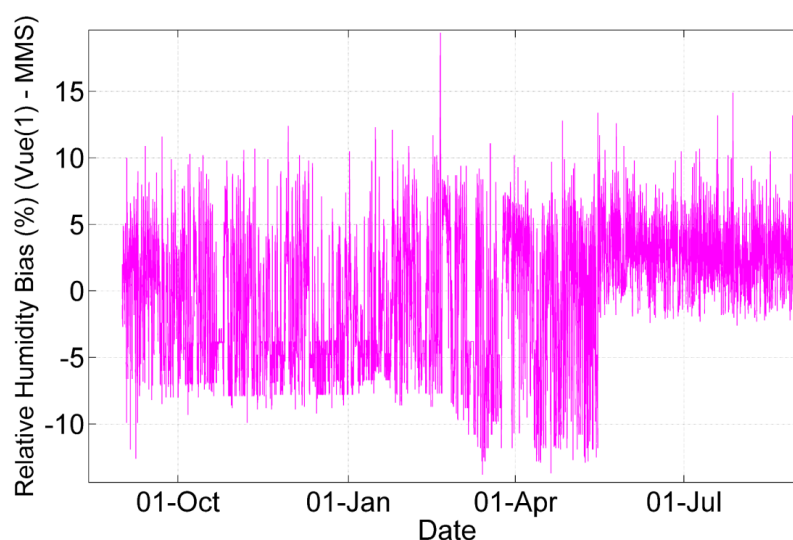


Figure 3.13. Time series of the Vue(1) station's relative-humidity bias, that is Vue(1) humidity – MMS humidity.

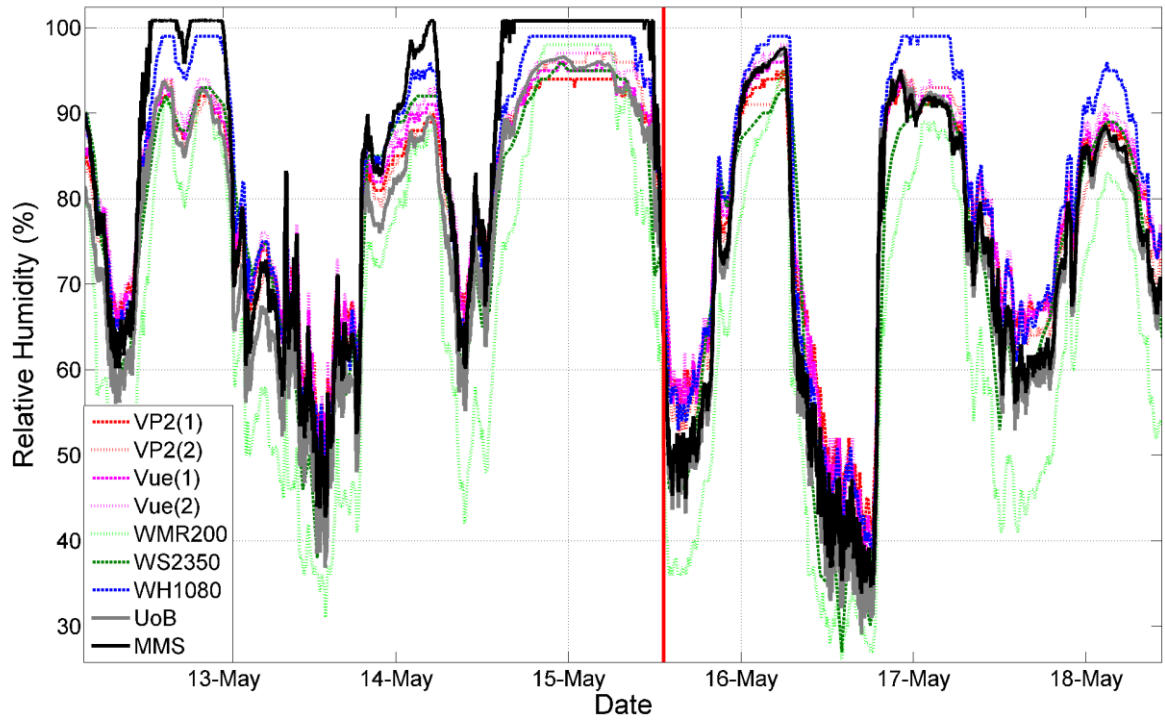


Figure 3.14. Relative humidity time series. Covers the period when the MMS's Rotronic HygroClip was changed, as indicated by the red line. Note the addition of the University of Birmingham (UoB) Vaisala humidity observations.

Data from the humidity sensors of the seven CWS we tested have very different patterns to their bias. The Davis VP2(1), Vue(1) and Vue(2) stations all show a wet bias over the majority of the humidity range (Figure 3.15). For all three stations the mean bias is greater than 3% under drier conditions (less than 90%), but when the humidity is greater than 90% the VP2(1) and Vue(1) exhibit small dry biases. These findings agree well with those of Burt during his review of the VP2 (Burt, 2009) and Vue model (Burt, 2013). The VP2(2) behaves slightly differently: under drier conditions it does not overread to the same degree as the other Davis stations, but it underreads more during wet conditions, with a greater residual variance across the whole humidity range.

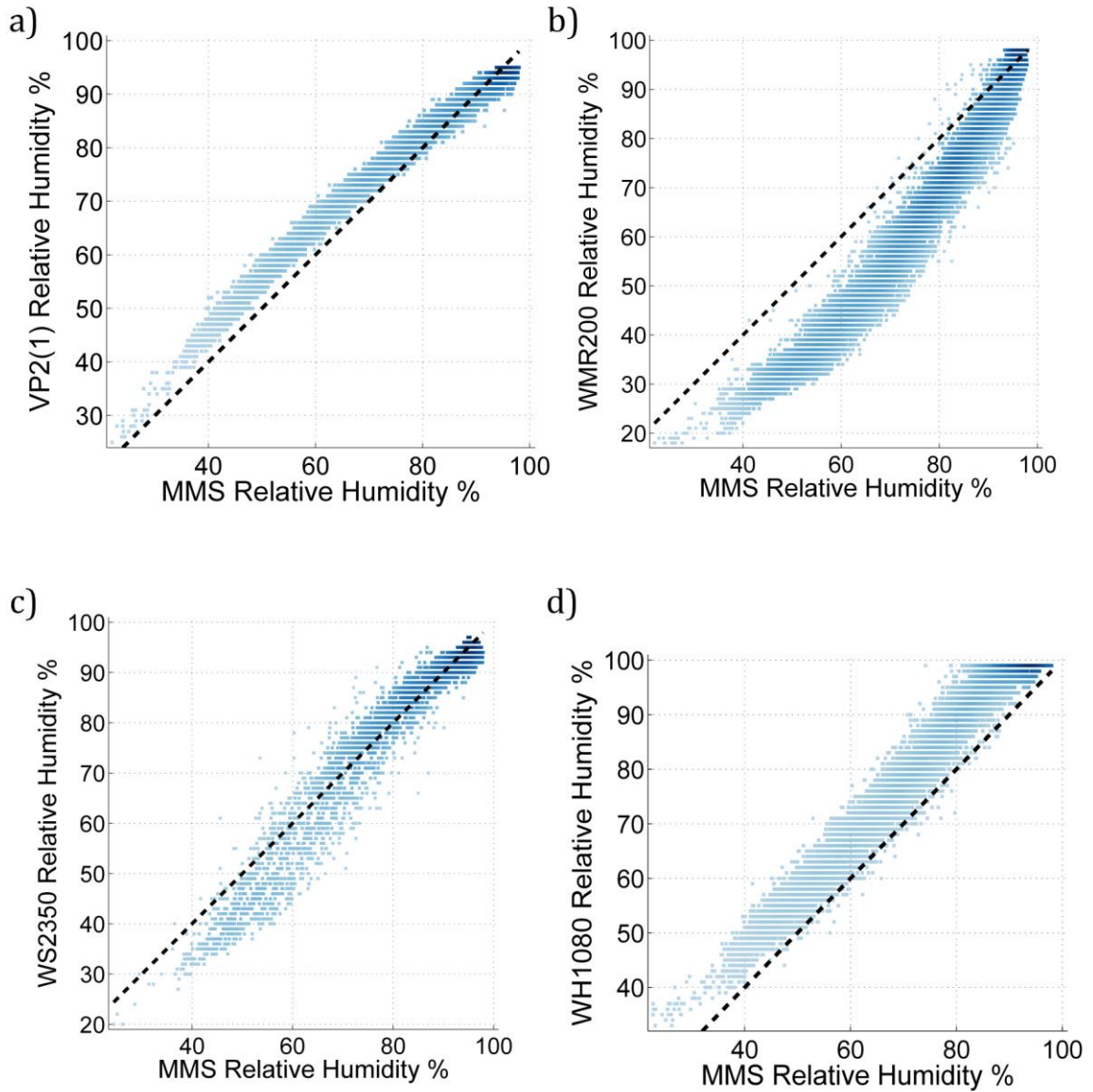


Figure 3.15. The CWS' versus MMS's relative humidity: (a) VP2(1), (b) WMR200, (c) WS2350 and (d) WH1080 stations. All observations between the 16th May 2013 and 31st Aug 2013 are shown. The darker the colour the greater the density of points.

The WMR200 underreads across the entire humidity range, and dramatically so in drier situations, where it exhibits a mean bias of -12.8%. The WS2350 also tends towards a dry bias during drier situations, but with a less extreme mean of -1.4%. Between 70 and 90% (Figure 3.15) this switches to a wet bias. By contrast, the WH1080 has a large wet bias over the entire humidity range, with an overall bias of 7.5%.

In this particular field study it was difficult to accurately quantify the response times of the different humidity sensors. We therefore suggest a further study, in which the humidity sensors (still inside of their radiation shields) are placed within a climate chamber. The chamber can initiate a step change humidity from which the response time of the different sensors can be derived with much greater accuracy.

As we anticipated some interaction of temperature and relative-humidity biases, we also considered dew-point temperature. The mean dew-point biases for all Davis stations were within 1 °C of the MMS. In agreement with other studies that found the VP2 station monthly means mostly within 1 °C of the reference sensor (Burt, 2009) and Vue station readings that were approximately 1 °C too high (Burt, 2013), both Vue stations in this study had a mean bias of +0.8 °C. The WS2350, with a mean bias of 0.9 °C was also within 1 °C of the MMS's sensor, but with a larger residual variance. The mean bias of the WMR200 and WH1080 stations was more significant, at -1.8 °C and 2.3 °C respectively (Figure 3.16).

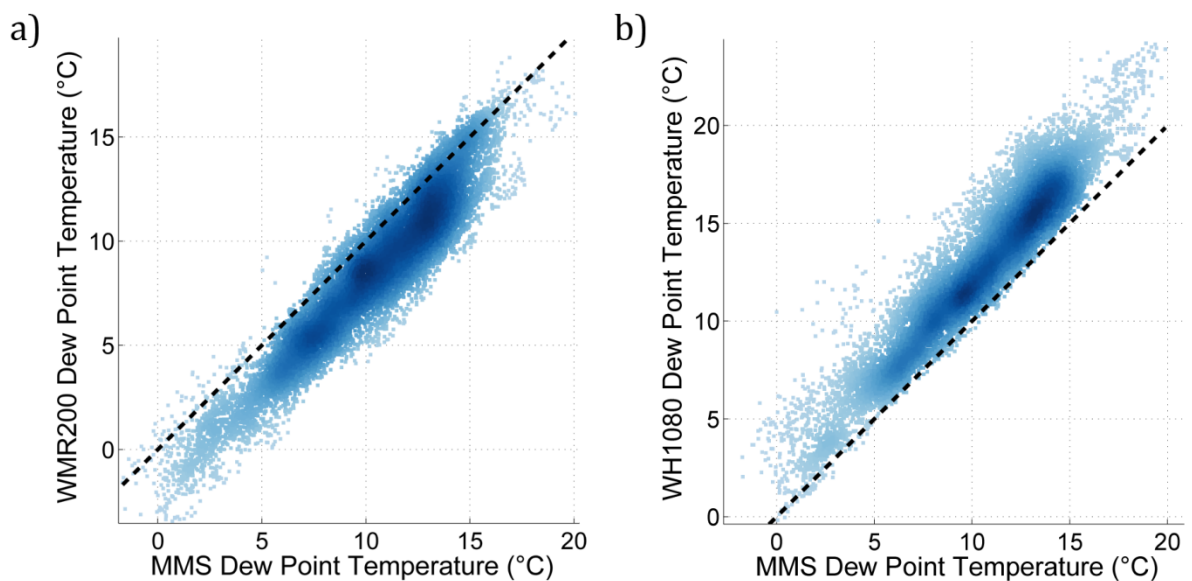


Figure 3.16. The CWS' versus MMS's dew-point temperature: (a) WMR200 and (b) WH1080 stations. All observations between the 16th May 2013 and 31st Aug 2013 are shown. The darker the colour the greater the density of points. The equivalent plots for the other CWS tested are shown in Appendix 8.7.

The dew-point values of the CWS were derived by the station's electronic console, however, the difference due to their use of potentially different algorithms was virtually negligible, never exceeding 0.02 °C.

Parameterising biases in relative humidity is a challenging task. In general there are two main sources of bias. First, there is the capacitive sensor itself. Figure 3.15 demonstrated that in comparison to the MMS's sensor the CWS have potentially large calibration errors. The magnitude and even the sign of the bias often changes depending on the humidity. Bias may also be induced from hysteresis, when the sensor's response to a change in humidity varies depending on whether the humidity is rising or falling. As with the MMS's Rotronics HygroClip, these CWS' sensors may drift over time, potentially becoming more biased the longer they are in the field. The

second source of bias comes from the inadequate shielding or housing of the sensor, and is closely related to the shielding problems that lead to temperature biases. For example if the shielding overheats, the air within the shield warms, reducing its relative humidity, thus causing an apparent dry bias. Alternatively, if humidity is falling after a period of saturated conditions, a poorly ventilated shield may prolong the time a sensor reads saturated. Trying to tease apart the two sources of error so that they can be parameterised is very difficult.

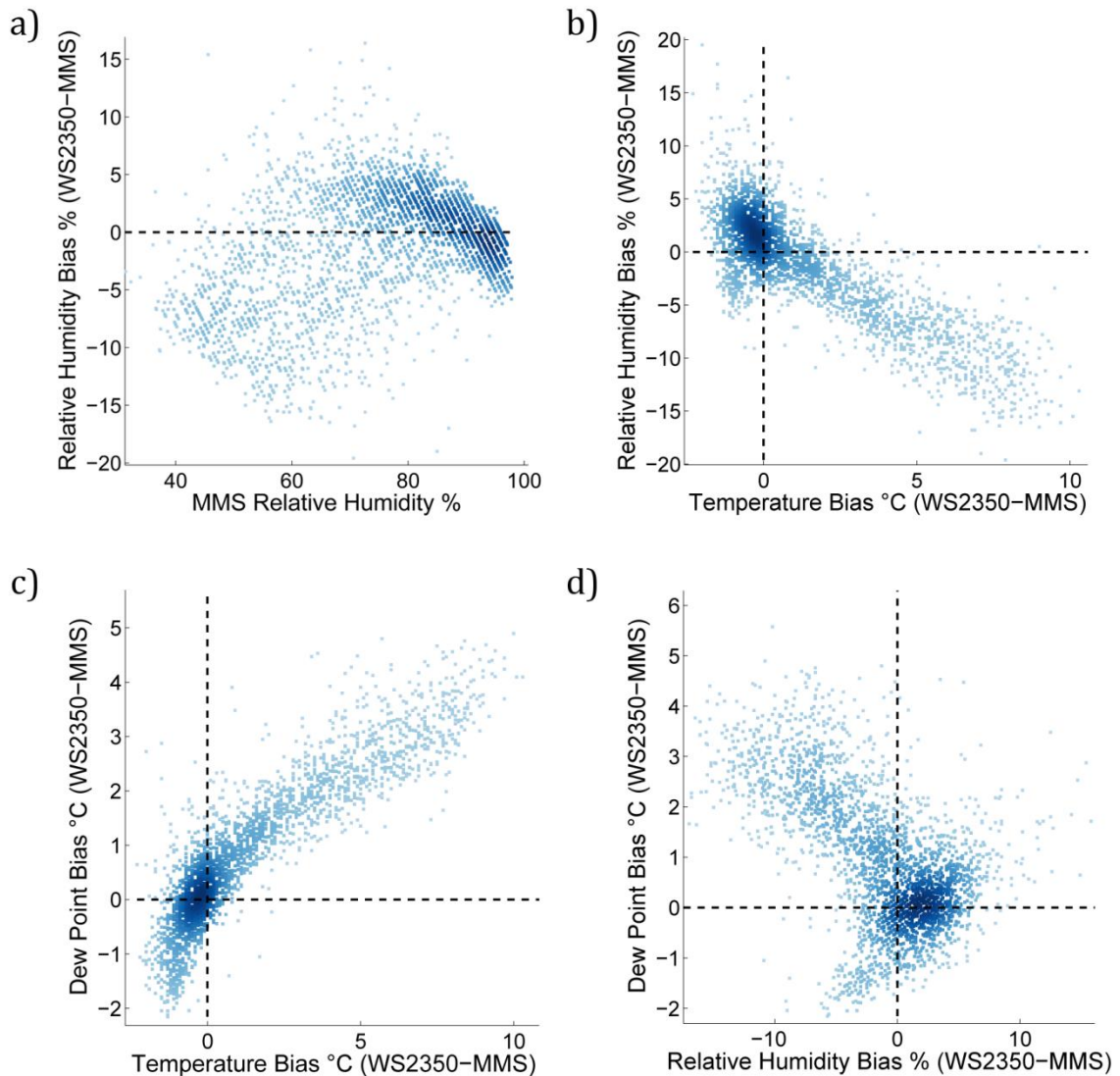


Figure 3.17. Plots of the relationship between temperature, humidity and dew point, and their biases for the La Crosse WS2350 station. All observations between the 16th May 2013 and 31st Aug 2013 are shown. The darker the colour the higher the density of points.

Figure 3.17 shows a series of plots for the WS2350 station that can help us decipher the source of its humidity and dew point bias. It is apparent in Figure 3.17(a) that the relationship between the humidity bias and the MMS's humidity is somewhat unclear. Figure 3.17(b) plots the humidity bias against temperature bias. The majority of points display a slight cold and wet bias, however, there is a long tail where warm

temperature biases are associated with a dry humidity bias. This may be caused by increased temperatures within the sensor housing, relative to the surrounding air. Bias in the dew-point temperature is inherited from both the humidity bias and the temperature bias. Unsurprisingly Figure 3.17(c) shows that a warm temperature bias leads to a subsequent warm dew-point bias, however, the relationship is not perfectly linear. As expected the dew point is too high when the humidity is wet-biased. However, it is also too high when humidity is dry-biased, perhaps because a warm temperature bias has not only caused a subsequent high dew-point bias but also a dry humidity bias when the sensor housing overheated. It is worth noting that we have chosen to plot just one station here: the plots for the other stations can look very different, further indicating that the parameterisations must be learnt for each individual CWS.

This field study has reinforced that relative humidity is a difficult variable to measure, for CWS and MMS stations alike. Having seen biases that slowly drift and suddenly jump within the reference MMS data it is crucial that before using the MMS network to correct real CWS observations that it too it undergoes its own quality control procedure. A network of more accurate sensors, such as chilled mirror hygrometers that measure dew-point temperature directly, could help anchor the MMS network and in turn the CWS network, although such a network may prove difficult to keep in good operational condition. Incorporating relative humidity into the bias correction model detailed in Section 5.6 would prove difficult. With an upper limit capped at 100%, it can produce non-Gaussian errors. Converting relative humidity into other variables that represent air moisture, such as dew-point temperature, may provide a solution to this. We have shown examples where relative humidity and dew-point biases are dependent on temperature bias, so it is important that we model the moisture variable jointly with temperature and its bias.

3.2.4. Rainfall

Figure 3.18 shows a plot of cumulative rainfall throughout the year-long study. All but the VP2(2) station measured totals less than the MMS's gauge. It is interesting that one VP2 station overread whereas the other underread, particularly as the VP2 model allows for calibration of the tipping buckets using a screw under each bucket. Before installing the VP2 stations they were calibrated in the laboratory so that on average both read within 2% of the truth. It is curious that, once in the field, they should deviate from the professional gauge by approximately 10%, and in different directions. When Burt (2009) tested a different VP2 station against a standard 'five-

inch' gauge he found the annual total was just 1.8% higher, but the agreement was not consistent, with monthly differences ranging from -10% to +19%.

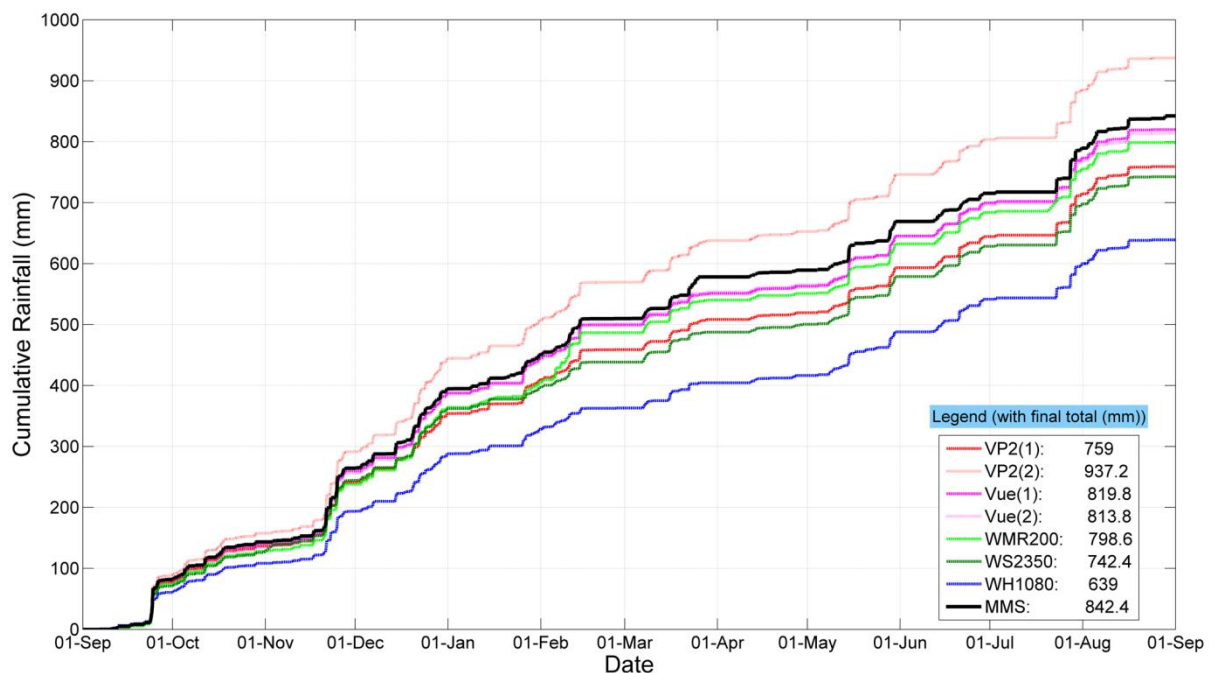


Figure 3.18. Cumulative rainfall totals of the seven CWS throughout the year-long study. Data from the professional Met Office gauge are shown by the black line. The final totals are displayed in the legend.

The Davis Vue stations show a very good agreement with the MMS's gauge and with each other, both undercatching by less than 4%. However, results by Burt (2013) show that this slight undercatch is not consistent throughout all Vue stations: as compared with the standard 'five-inch' gauge the annual total of their Vue stations was 9% too high. The WMR200 showed a reasonable agreement, undercatching by just 5.2% at the end of the period. The WS2350 and the WH1080 stations undercatch by greater amounts, with a yearly total of just 88% and 76% of the Met Office total respectively.

Six of the seven CWS tested displayed yearly rainfall totals lower than the MMS's rain gauge. One explanation for this could have been that, at 1–2 m above the ground, the CWS' rain gauges are mounted higher than the MMS's gauge at 30 cm, as such they experience higher wind speeds, which can lead to undercatch (Guo, et al., 2001). The shape of the gauge may also have an impact on these wind-induced errors (Sevruk & Nespor, 1994). In this study we looked for a relationship between wind speed and the daily CWS' rainfall totals as a proportion of the daily MMS's total. However, no obvious relationship was found, either for average wind speed taken at times of measured rainfall, or for daily average wind speed. Measurements from both the CWS's anemometer and the Vector Instruments anemometer, mounted at 7m, were used separately. Possible explanations for this lack of relationship is that the

Winterbourne No. 2 site is relatively sheltered, with a mean 7 m wind speed of 1.6 m s^{-1} and a maximum of 10.2 m s^{-1} . The anemometers of the four Davis instruments never read any higher than 5 m s^{-1} . It is possible that these relatively low wind speeds did not cause noticeable undercatch and that the biases seen had a different source. The MMS's gauge is deep with steep sides to prevent heavy rain from bouncing out, while all CWS tested, bar the VP2s, are much shallower, particularly the WS2350 and WH1080. Such a design makes them prone to undercatch due to rain drops bouncing out, an effect that potentially outweighs any influence of wind speeds. Wind direction may also play a role with the height and shape of upwind obstacles varying with direction. However this effect was not looked at in detail, and remains an area for further investigation.

Even before deploying the CWS' rain gauges outdoors the tipping buckets within may be poorly calibrated, producing a bias straight out of the box. To test this, 500ml of water was slowly dripped through each rain gauge indoors. By dividing the volume of water by the area of the gauge it is possible to calculate the depth of rain (in mm) that the station's console should display (Overton, 2007). Differences between expected and measured depth were as large as 13%, with most stations underreading. A corresponding correction was then applied to the yearly cumulative rainfall totals. For some stations this led to a small improvement, but for others their yearly total was made much worse. Clearly other factors are at work outdoors that outweigh errors due to poor calibration, proven by the poor performance of the VP2 stations in the field despite being calibrated to less than 2% error indoors.

Some CWS' rainfall bias, particularly when dealing with daily totals, can result from cold and snowy synoptic conditions. The gauges of the CWS tested were prone to filling with snow (e.g. Figure 3.19) when the MMS's gauge did not. This resulted in delayed tips when the snow finally melted. The funnel exit hole of the CWS' gauges was also prone to freezing over, again causing a delay in rainfall readings. The data could be split into cold and warm periods to better demonstrate the impact of these cold weather effects; however this remains an area for further investigation.

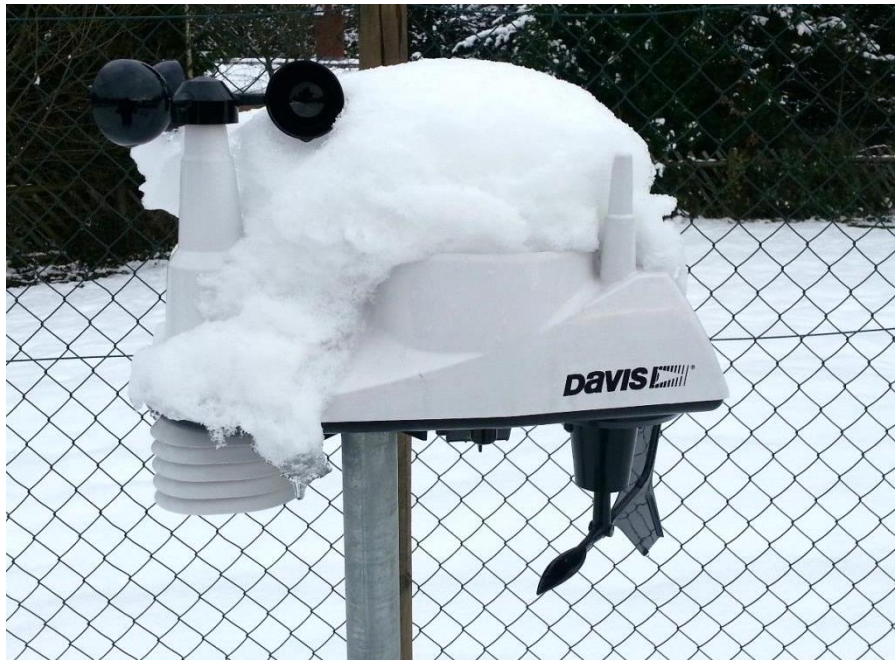


Figure 3.19. Davis Vantage Vue on 24th January 2013. Note that the rain gauge located on top of the unit is completely filled with snow, so much so that it prevents the wind cups from fully rotating.

The CWS are usually good at capturing the intensity and timing of rainfall events, but their long-term cumulative total can differ from professional gauge measurements significantly. Using the MMS's gauge as our best estimate of the truth, Figure 3.20 shows that by identifying the relationship between the MMS's and CWS' cumulative rainfall time series we can correct the CWS' time series to fall in line with that of the MMS. This works well because, for the CWS tested here, the proportion of undercatch or overcatch tended to remain relatively constant through time, thereby allowing for a correction to be learnt from preceding data. At each timestep the following steps are taken in order to make this correction:

1. Leading up to a given timestep there is a time series of cumulative rainfall totals from both the MMS and CWS gauges. Any preceding timesteps where the CWS total does not change (i.e. dry spells) are filtered out of the dataset.
2. A simple linear scaling correction is then derived from the remaining pairs of MMS and CWS cumulative totals, assuming a fixed origin at (0,0). For example if a CWS consistently undercatches by 10% then the calculated correction will be 0.9.
3. The CWS cumulative rainfall total at the current timestep is then corrected by dividing its uncorrected total by the calculated correction.

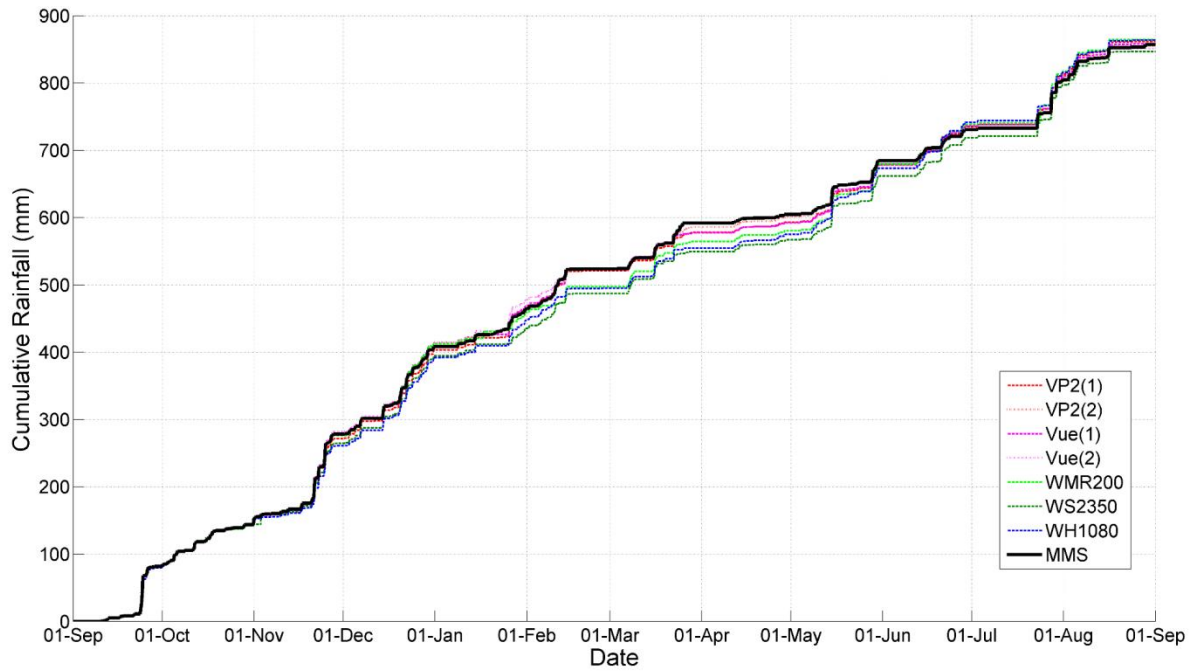


Figure 3.20. Plot of cumulative rainfall totals from the MMS’s and seven CWS’ gauges. The cumulative rainfall values (available every 10 min) for all CWS were corrected using the relationship between their cumulative rainfall and that of the MMS’s gauge; learnt from preceding data only.

In reality, a professional gauge is rarely collocated alongside a CWS, in which case observations must be carefully interpolated from nearby professional gauges, for example, from the MMS or Environmental Agency networks. Effectively merging radar accumulation data with this gauge data could improve the interpolation (DeGaetano & Wilks, 2009), which should improve CWS’ long-term totals, while keeping the detail of individual and isolated rainfall events captured by the CWS.

3.3. Summary

This intercomparison field study has crucially shown it is possible to identify, parameterise, and correct biases evident within CWS data, and the size of the observed biases highlights just how important such corrections are.

In particular we detailed the significant radiation-induced temperature biases, and how weighted global radiation observations made at MMS sites can be used to parameterise and correct them. In theory this should allow us to correct operational CWS data, given a reliable estimate of the radiation at CWS locations. Producing radiation estimates at operational CWS locations requires a radiation interpolation model as described in Section 5.3.3; a model that leverages the lagged nature of the relationship. We also noted the impact that micro-scale siting can have on the strength of incoming radiation through shading effects. Modelling this effect is very difficult,

however in Section 5.5.1 we propose an approach that can remotely quantifying the degree of sheltering at a CWS location.

We also saw that different models of station exhibit dramatically different bias magnitudes and relationships. To correct real CWS data it is therefore crucial to decipher which model is being used in order to apply the right correction. An approach to do this is discussed in Section 5.4, highlighting the importance of metadata. The temperature–radiation relationships observed in this study for each design of radiation shielding are used to inform the corrections applied to operational CWS of the same design (Section 5.6.3).

There were also examples where stations of the same model actually showed very different biases; take for example the rainfall biases of the two VP2s. Therefore, although knowing the model type gives us some *a priori* information about the types of biases we would expect, this belief should be updated if the data itself indicates otherwise. However, unlike this field study, operational CWS are not be collocated with professional instruments from which we can easily quantify the instrumental biases. To overcome this problem an interpolation model is described in the next chapter (Chapter 4) capable of interpolating temperature observations from the professional MMS network to the CWS locations.

The CWS tested also displayed some calibration biases. For example the Vue(2) station appeared to display a consistent cold bias of around $-0.2\text{ }^{\circ}\text{C}$ (masked by longwave radiation-induced biases during the day, Figure 3.2). Section 5.6 details an approach for learning these temporally-consistent biases over time.

4. Interpolating professional observations

In order to learn whether a given CWS displays any biases, it is crucial that we have a reliable independent estimate of the weather at that CWS's location. As a CWS may be tens of kilometres from the nearest professional MMS station we rely on an interpolation model to interpolate professional observations to CWS locations. In this chapter we detail the interpolation model that was used and test its performance using cross-validation with professional MMS temperature observations over four case study periods. As previously mentioned we focus only on interpolating air temperature. The interpolation of dew point and precipitation is a topic for further work.

By comparing the interpolated professional observations against the CWS data, we attempt to gradually update our expectation of the CWS's bias over time. In Chapter 5 we discuss exactly how this is done. If the bias is successfully learnt, then once it is removed (assuming negligible model and representativity errors) the only difference relative to the interpolation model's prediction should be natural spatial variations that the interpolation model was unable to resolve. This is where the value in the CWS data lies. As estimating the bias is a difficult task, we also aim to accurately quantify our uncertainty, in order to express our confidence in the learnt bias correction.

The task of interpolating land surface meteorological observations is one that has been performed numerous times to date, with a variety of techniques being implemented and compared (e.g. Daly (2006); Hofstra, et al., (2008)). Approaches include thin plate splines (Hijmans, et al., 2005), various forms of kriging (Hengl & Heuvelink, 2007), local regression (Daly, et al., 2002) and artificial neural networks (Rigol, et al., 2001). Some studies use forecast model initialisations to help guide the spatial interpolation (Degaetano & Belcher, 2007), and others stratify their interpolation models by circulation pattern (Courault & Monestiez, 1999).

Much of the literature focuses on the interpolation of irregularly spaced climatological station data, such as daily mean, minimum and maximum temperatures, to a regular grid. The key difference in this study is that we use instantaneous sub-daily temperature measurements (i.e. valid for a given minute) and although our approach could be used to interpolate to a regular grid, here it is only necessary to interpolate to the locations of CWS. Our interpolation model is run over a domain that consists primarily of Great Britain, but also some surrounding islands (Figure 2.2). By using predictors (Section 4.4) with a high spatial resolution, e.g. forecast model data on a 1.5

km resolution grid, and land cover maps at 25 m resolution the aim is to produce an estimate as close to ‘true’ temperature at a CWS location as possible. However as with all complex models this model is not perfect. The resolution of our predictor data is finite, other important predictors may be unresolved, and errors may be introduced by an imperfect model structure. As such the scale to which our predictions are valid is probably limited to a few kilometres around the CWS location. The exact scale links to the concept of representativity (Section 2.3.5). In the bias correction model (Section 5.6) a specific term is introduced to handle such representativity errors. However as long as our estimates remain unbiased on average then over time CWS biases may be corrected for (as detailed in Section 5.6) and the CWS’s own observations thus provide the finer resolution estimate of the temperature at its location.

As detailed below (Section 4.4.1) our interpolation model incorporates short range forecasts from a numerical weather prediction model. As a result our model borrows several attributes common to the post-processing procedures used to handle the output from such models (Moseley, 2011). For example, bilinear interpolation is used to map the gridded forecast model output to station locations, before applying a height correction to account for differences in model cell height and station elevation. We also use real professional observations to learn the biases in the forecast - biases that are either spatially correlated, or correlated to one of the other predictors. While post-processing systems learn these forecast model biases with the aim of correcting their forecasts of upcoming weather, we instead use past forecasts and ‘correct their biases’ in order to best model the spatial temperature field.

4.1. Input MMS data

To run and test our interpolation model, professional temperature observations from the aforementioned Met Office MMS network are used (Green, 2010). The MMS network consists of over 200 land surface stations (Figure 2.2) that meet strict WMO quality standards (WMO, 2010). It is important to mention that MMS temperature observations are stored by the Met Office at hourly and minute resolutions. Here we only use hourly resolution temperature data (although in Section 5.3.3, minute-resolution global radiation observations are required). There are pros and cons to using this hourly resolution data over the minute data. The advantage is that this hourly data has been fed through the Met Office’s quality control system, which flags up erroneous data which we then discard from our dataset. The disadvantage is that these hourly observations are in fact valid at 10 minutes *to* the hour rather than *on* the hour. This is an artefact of manual observation practises, where the 10 minutes allowed observers to collect, code, and submit their observations before the deadline

on the hour. As with the MMS data from the field study (Section 3.1.1), the minute observation used at 10 to the hour is the mean of four 15 s samples taken over the minute. As the interpolated MMS observations are used to learn biases in CWS observations it is important that the CWS observations are also valid at, or close to, 10 minutes to the hour (discussed further in Section 5.1). The unavoidable issue here is that the Met Office's short range forecast, used as a predictor within this interpolation model (Section 4.4.1), is valid *on* the hour. We therefore have to assume that the impact of this difference is negligible and that the design of the interpolation model is such that it can compensate for this.

4.2. Case study periods

In order to thoroughly evaluate the performance of our interpolation model, it is important to test it over a range of meteorological conditions. To this end, we selected four 2 week periods, one from each season, which included a range of different synoptic situations. Below is a summary of the synoptic situation during each period. In Chapter 5 these same periods are used to test our bias correction model using real citizen data gathered over each period.

4.2.1. Autumn

Period: 1st – 14th October 2012

Summary: A mainly cold and unsettled period with frequent rain. However, a weak ridge between the 6th and 10th brought drier and sunnier weather, and also night frosts. The first five days consisted mainly of sharp showers and sunny intervals, although heavy and prolonged rain fell across southern England and Wales from late on the 4th through to early on the 6th (Eden, 2012).



Figure 4.1. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Autumn period: 1st – 14th October 2012. Synoptic charts from www.wetterzentrale.de archive.

4.2.2. Winter

Period: 17th – 30th January 2013

Summary: This winter period, characterised by cold, cloudy, and most notably snowy conditions, was selected as a challenging period. For the first 9 days high pressure over Scandinavia led to several snowfalls, notably on the 18th, 20th, 21st and 22nd. On the 25th an occlusion from the Atlantic brought about heavy snowfall before heralding a milder, but wet and windy, end to the period. Light winds before the 25th were replaced with strong gusts particularly for the 28th-30th (Eden, 2013a).



Figure 4.2. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Winter period: 17th – 30th January 2013. Synoptic charts from www.wetterzentrale.de archive.

4.2.3. Spring

Period: 13th – 26th May 2013

Summary: A cold May with some heavy frontal rainfall. The period began with an unsettled, cold and often windy cyclonic/westerly regime. A deep depression crossed the UK on the 14th and 15th bringing heavy banded rainfall with strong winds following up behind it. A fine bank holiday weekend, 25th-26th, rounded off the period (Eden, 2013b).



Figure 4.3. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Spring period: 13th – 26th May 2013. Synoptic charts from www.wetterzentrale.de archive.

4.2.4. Summer

Period: 24th June – 7th July 2013

Summary: A mostly dry and fine month, selected to highlight CWS radiation biases. The period began dry with occasional sunshine and near-normal temperatures, but light rain fell on the 27th and 28th before turning fine and warm for final two days of the month (Eden, 2013c). July began cooler and changeable, but from the 4th an anticyclonic situation took over, leading to dry, sunny and very warm conditions for nearly all of the British Isles (Eden, 2013d).



Figure 4.4. Met Office surface pressure charts valid at 00:00 GMT at the start of each day. Summer period: 24th June – 7th July 2013. Synoptic charts from www.wetterzentrale.de archive.

4.3. Interpolation model design

The interpolation model's design is based upon a Bayesian linear regression model (Equation (1)). At each timestep global regression is performed – using the training data to learn the distribution of the regression coefficients, β , in order to predict the temperature at the test sites. 'Global' is used in the sense that the regression coefficients apply across the entire British domain. In verification mode, as used in

this chapter, data for the MMS sites form both the training and test data through cross-validation. Operationally, and as used in Chapter 5, all the MMS sites are used to train the model in order to predict at CWS locations. The Bayesian framework, detailed below (Section 4.3.2), ensures that these regression coefficients are not learnt from scratch at each timestep. They propagate through time – informed both by what was learnt at preceding timesteps, i.e. the ‘prior’, and by the addition of new data to yield a posterior belief. This ensures that the model remains stable through time. This is particularly useful if it is necessary to deal with small sample sizes.

The predictors that form the basis functions within the design matrix X (Equation (1)) include a short range forecast from the UKV numerical weather prediction model, various geophysical properties characterising each station, along with Radial Basis Functions (RBFs) in order to provide some localisation in an otherwise global model. These basis functions are explained in more detail in Section 4.4.

4.3.1. Linear regression

Here we outline the basic structure of the Bayesian linear regression model used. For more details see Gelman, et al., (2003). Appendix 8.1 details the equation notation used here and in the rest of the thesis. The prediction is given by:

$$y = \beta^T X + \epsilon, \quad (1)$$

where

$$\begin{aligned} \epsilon &\sim \mathcal{N}(0, v_\epsilon), \\ v_\epsilon &\sim \mathcal{IG}(a, b), \\ \beta|v_\epsilon &\sim \mathcal{N}(\mu_\beta, v_\epsilon \cdot \Sigma_\beta). \end{aligned}$$

Equation (1) makes a prediction, y , of the ‘true temperature’ at the test station locations with a predicted error ϵ . X is typically called the design matrix and contains our basis functions (Section 4.4). β is the regression coefficient parameter vector learnt from the training data. Note the T , which indicates that the transpose of the β matrix is used. Say, for example, that X was simply a vector of elevation values at the test locations, then β would simply represent the effective lapse rate learnt from the training data with a given mean, μ_β , and variance, $v_\epsilon \cdot \Sigma_\beta$. By default X always contain a constant term. Although the model is linear in parameters the design matrix can contain non-linear basis functions. A key benefit of using such simple models is their speed and interpretability (Bishop, 2007).

The model uncertainty, v_ϵ , becomes the learnt discrepancy between the model predictions and the observations and therefore incorporates both observation and model error. We use a Normal-inverse Gamma distribution jointly over v_ϵ and β . These are coupled and conjugate which provides closed form updates for the posteriors, meaning that the code can be run quickly and efficiently. The joint distribution ensures that as our model uncertainty, v_ϵ , increases so does our uncertainty about the regression coefficients.

This structure suits variables, such as air temperature, which when interpolated will display Gaussian interpolation errors. It is therefore unsuitable for variables, such as precipitation, that are likely to display non-Gaussian errors.

4.3.2. Bayesian framework

At each timestep three key steps occur as part of our Bayesian approach: forecast, update, and prediction at test/CWS locations; each step is explained in detail below.

Forecast

In order to use the posterior distributions of the regression coefficients and model uncertainty from the last timestep as priors at this timestep we must propagate them forward in time. Note that although the mean estimate of the regression coefficients remains the same (Equation (2)) we inflate our uncertainty about the regression coefficients (Equation (3)) and our model uncertainty (Equations (4) and (5)). By inflating the uncertainty the model is able to react to changes in the relationship between the predictors and the predictand without becoming over confident in its estimates.

$$\tilde{\mu}_{\beta,t} = \mu_{\beta,t-1} \quad (2)$$

$$\tilde{\Sigma}_{\beta,t} = \Sigma_{\beta,t-1} + \left(\Sigma_{\beta,t=0} \frac{\delta t^2}{\gamma_\beta} \right) \quad (3)$$

Here δt is the time between the last timestep and the current timestep, and should be in the same units as γ_β , the forgetting rate parameter for the regression coefficients. $\Sigma_{\beta,t=0}$ is the covariance matrix of the regression coefficients at timestep zero, i.e. when the model was initiated; it is used to scale the added uncertainty. The ratio of δt to γ_β controls the rate at which the previous timestep's regression coefficient information is

forgotten. Here γ_β is set as 24 hours, i.e. if the timestep between model runs, δt is greater than 24 hours the regression coefficients learnt at the previous timestep become at least as uncertain as before any data was seen. In this study however δt is always 3 hours due to the model's reliance on the short range forecast model (Section 4.4.1). Because the model receives a lot of new data at each timestep it proves fairly insensitive to the value of γ_β .

To scale the variance of the residuals from the model we also inflate the Gamma component:

$$\tilde{a}_t = a_{t-1} \quad (4)$$

$$\tilde{b}_t = \frac{b_{t-1}}{2} \quad (5)$$

The denominator in Equation (5), i.e. 2, was set empirically.

Update

The update step combines the ‘prior’ distributions of the regression coefficients and the model uncertainty with the new data made available at this timestep to produce posterior distributions from which predictions can be made.

In Equation (6) the regression coefficients’ covariance matrix $\tilde{\Sigma}_{\beta,t}$, handled as a precision matrix $\Sigma_{\beta,t}^{-1}$ is updated using the new design matrix X :

$$\Sigma_{\beta,t}^{-1} = X^\top X + \tilde{\Sigma}_{\beta,t}^{-1} + \Gamma, \quad (6)$$

where Γ is a regularisation term, a diagonal matrix with the value 0.000001 along the diagonal used to ensure numerical stability.

The mean estimate of regression coefficients, $\tilde{\mu}_{\beta,t}$, is also updated using the new design matrix along with the vector of target temperature observations, t (Equation (7)). Note how the updated regression coefficient covariance matrix, $\Sigma_{\beta,t}$, from Equation (6) is also used.

$$\mu_{\beta,t} = \Sigma_{\beta,t}(\tilde{\Sigma}_{\beta,t}^{-1} \tilde{\mu}_{\beta,t} + X^\top t) \quad (7)$$

Equations (8) and (9) denote the update of the model uncertainty parameters a and b :

$$a_t = \tilde{a}_t + 0.5m, \quad (8)$$

Where m is the number of training stations.

$$b_t = \tilde{b}_t + 0.5(t^\top t) + \tilde{\mu}_{\beta,t}^\top \tilde{\Sigma}_{\beta,t}^{-1} \tilde{\mu}_{\beta,t} - \mu_{\beta,t}^\top \Sigma_{\beta,t}^{-1} \mu_{\beta,t}. \quad (9)$$

Prediction at test/CWS locations

Equations (10) and (11) use the posterior distributions learnt from the update step to make temperature predictions at the test stations. These predictions are distributions with a mean estimate $\mu_{y,t}$ and covariance matrix $\Sigma_{y,t}$. This covariance matrix is crucial as it provides an estimate of the uncertainty of our prediction. In Section 4.5 we evaluate the accuracy of these mean predictions and assess how well the model performs probabilistically.

$$\mu_{y,t} = X \mu_{\beta,t} \quad (10)$$

$$\Sigma_{y,t} = \frac{b_t}{a_t - 1} (1 + (X \Sigma_{\beta,t} X^\top)) \quad (11)$$

The mean estimate $\mu_{y,t}$ and covariance matrix $\Sigma_{y,t}$ are the primary outputs from this interpolation model. In Chapter 5 (specifically Figure 5.32) we show how these two outputs enter our bias correction model.

4.3.3. Alternative clustered approach

The temperature interpolation model detailed above assumes that the regression coefficient, β , for each predictor (Section 4.4) is constant across the entire country at a given timestep. However it is likely that the regression coefficients not only vary temporally, as already captured by the model, but also spatially. Taking the regression coefficient for elevation as an example, we would expect that when Britain is covered by several different air masses the lapse rate probably varies between them. Below we detail a novel approach that was also tested in order to capture this effect. However as this approach was more computationally expensive and provided little improvement in the cross-validation error the approach already described above was adopted instead.

This alternative approach implemented a Gaussian Mixture Model to cluster professional and CWS stations alike into clusters. A MATLAB implementation of a Gaussian Mixture Model was provided by the Netlab toolbox (Nabney, 2002). The regression coefficients and model uncertainty could then be learnt and updated for each cluster individually (Figure 4.5). It is the addition of this clustering step, and the fact each cluster must now be processed individually that increases the computational time. Soft assignments were used so that each station belonged, with a given probability, to each of the clusters, although usually each station significantly favoured a particular cluster. For stations with a strong membership to a cluster their data was given a larger weighing than those with a weak membership when the regression coefficients and model uncertainties were updated for that particular cluster.

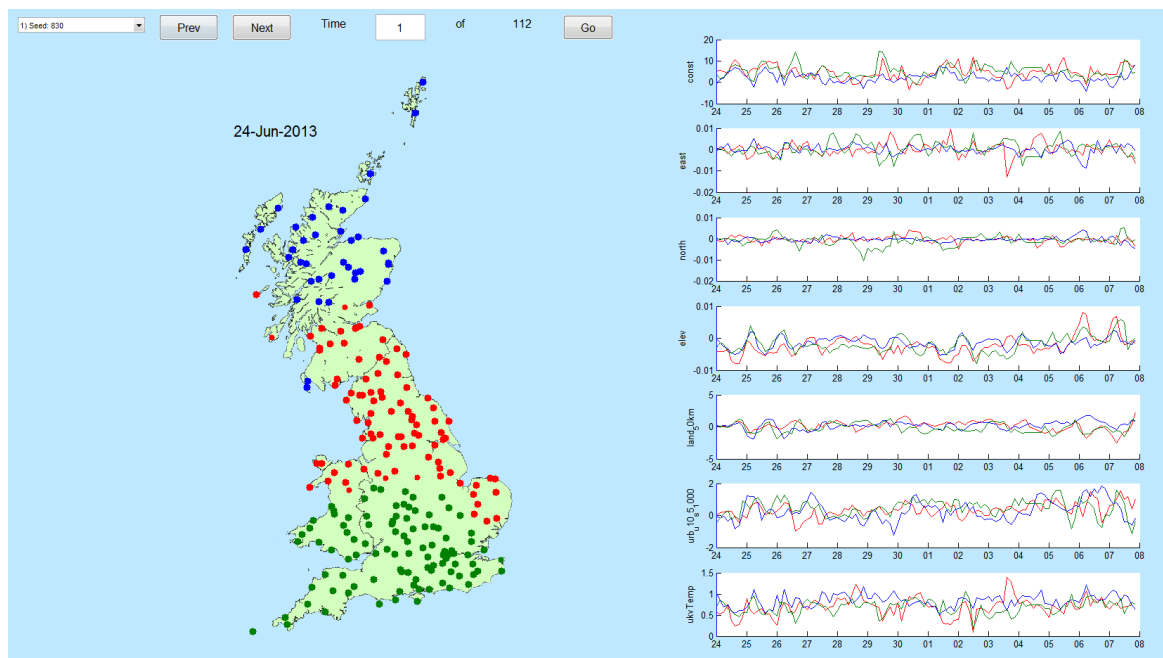


Figure 4.5. Screenshot of the graphical user interface (GUI) used to visualise the evolution of the clusters through time. The colour of each station represents the cluster to which its membership is strongest, shown here on the map (left) for a single timestep. The time series on the right of the plot shows the mean regression coefficient values for every basis function evolving over the 2 week summer period, at 3 hourly intervals. Each of the 3 clusters is represented by a different line/colour. Only MMS stations (225 total) were used here.

The clusters were assigned based upon observations (or estimates) of the sea-level temperature, humidity, cloud cover, and wind speed at each station's location. These variables help characterise the synoptic conditions across the country, ensuring that the clusters loosely represent the overlying air masses. From Figure 4.6 it is possible to see how these variables influence the final clusters. The geographic easting and

northing coordinates of the stations were also used as inputs to the clustering, to ensure that the clusters remained spatially coherent.

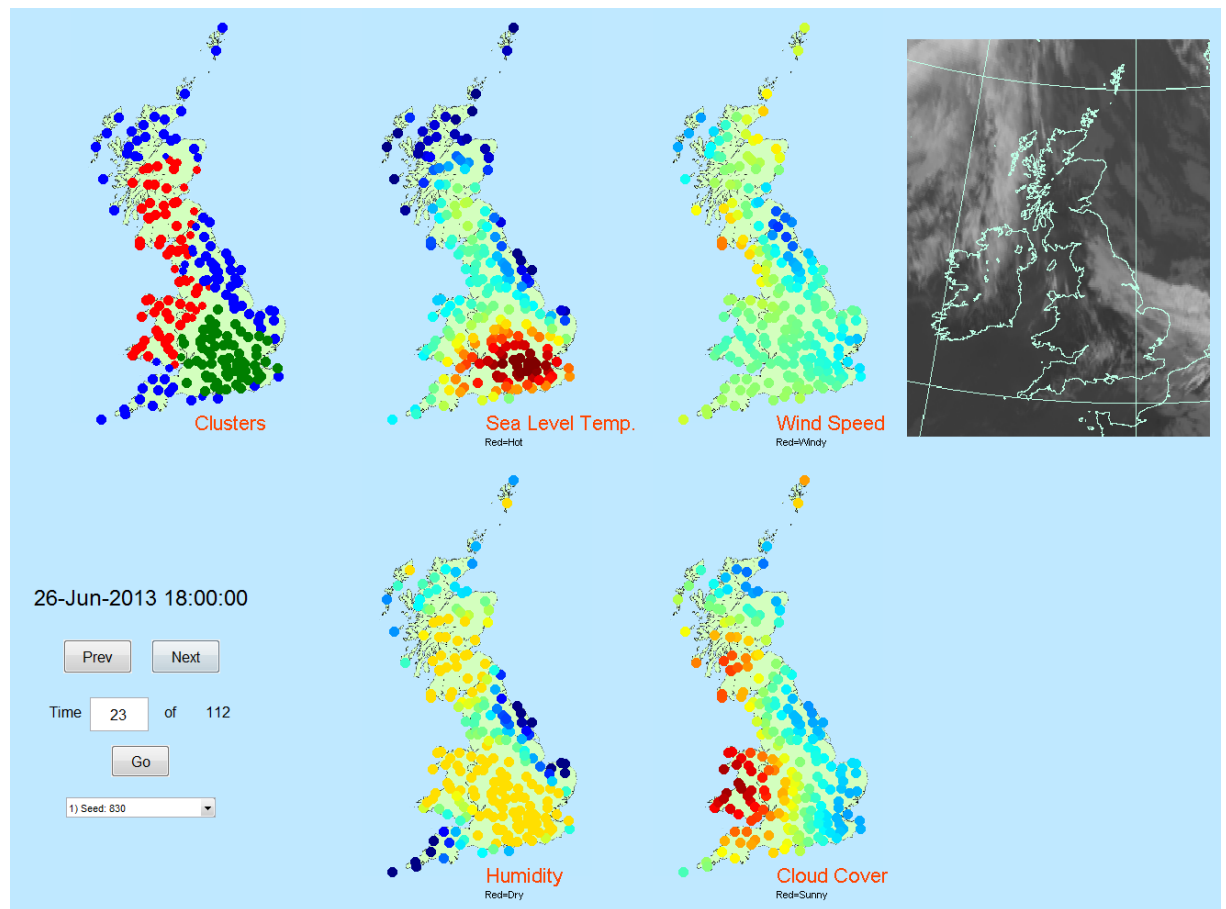


Figure 4.6. Screenshot of a GUI used to visualise the cluster assignments and how they correspond to the variables used to assign the clusters. The colour of each station in the *Clusters* plot denotes the cluster to which its membership is strongest (i.e. the cluster to which it has the greatest probability of belong to) and the size of each marker is indicative of the strength of its assignment to that cluster. Only MMS stations were used here.

These dynamic clusters evolve naturally through time (Figure 4.7). The number of clusters were fixed; with 3-5 clusters commonly used. It is very unlikely that Great Britain would be subject to more than 5 air masses at a given time; therefore using any more clusters would be unrealistic and could result in clusters to which very few stations are strongly assigned.

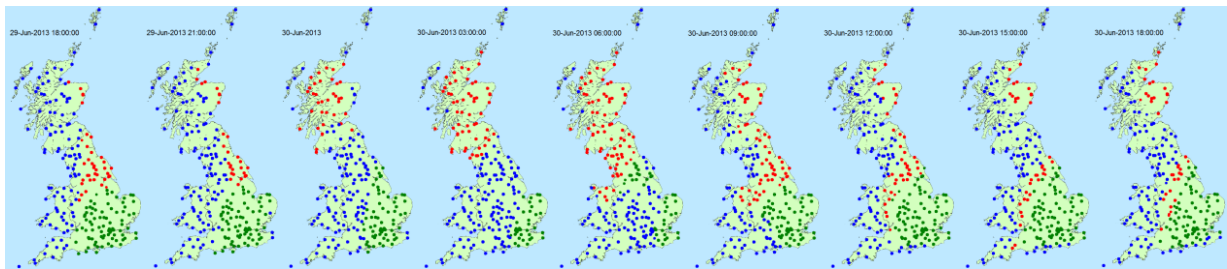


Figure 4.7. Snapshots of the clusters evolving through time, here at 3-hourly intervals. Colour denotes strongest cluster membership for each station. Only MMS stations are shown here.

Figure 4.8 illustrates the number of stations used to update the regression coefficients and model uncertainties for each cluster through time. Note that at times a cluster is informed by virtually no stations. This implies there are fewer air masses than clusters and therefore not all the clusters are required.

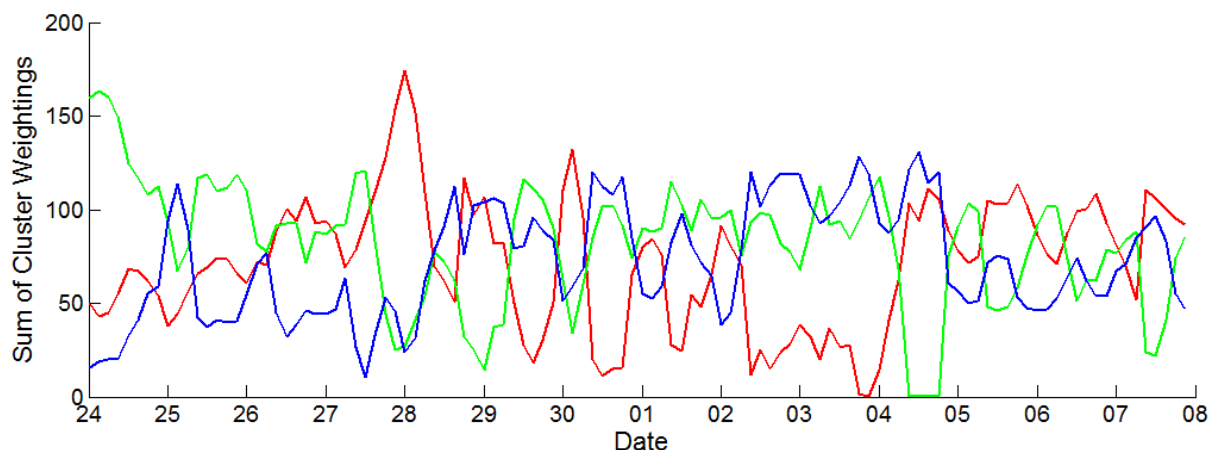


Figure 4.8. Time series of the sum of cluster weightings for each cluster, with each line representing a different cluster.

It was interesting that this clustered approach gave virtually no improvement over the chosen global approach. It is probable that the inclusion of the UKV numerical weather prediction model (Section 4.4.1), which inherently model different air masses, was able to capture this effect without the need for a clustered model. The use of Radial Basis Functions (Section 4.4.5) in the selected approach also helps resolve spatial variations. Further work would be required to show the potential benefit of such a clustered approach.

4.4. Basis Functions

The following sections (4.4.1 – 4.4.5) introduce the predictors that were chosen to form the basis functions within the design matrix X of our linear regression model. In deriving this final set of predictors many different predictors were tested, selected

because they were believed to be strong spatial covariates of surface temperature. Previous studies, such as Jarvis and Stuart (2001), found northing, elevation, coastal and urban effects to be particularly significant for the UK, although their study focused on explaining the variation in minimum and maximum daily temperatures. Unsurprisingly Alvarez, et al., (2014) also found that elevation was the leading covariate for maximum, mean and minimum daily temperatures in their study in the Western United States, with distance to coast also helping to reduce the Root-Mean-Square Error (RMSE). Each of these variables has been incorporated in some form below with the hope that they are also strong covariates of instantaneous sub-daily temperature, as used in this study.

As well as these geophysical predictors it was also important to incorporate a dynamic predictor, which unlike the geophysical predictors would vary through time. For this the Met Office's UKV model was used as explained below (Section 4.4.1). This would help capture the fluid nature of the near surface temperature field; resolving spatial variations the geospatial predictors alone would miss.

For each of these predictors only a 1st order basis function is used. 2nd order basis functions, e.g. x^2 , were also tested for each predictor individually, but provided little or no improvement to the overall accuracy of the model. They were therefore omitted to save on computational cost. Radial basis functions are also incorporated, as explained in Section 4.4.5. In Section 4.5.1 we demonstrate the value each of these basis functions adds to the interpolation model.

Each basis function requires an initial prior distribution of its corresponding regression coefficient to initialise the model. In order to sensibly specify the mean and variance of each regression coefficient we simulated a temperature field across the entire Great Britain domain (1 km resolution; 700×1300 cells, i.e. the same domain as the Ordnance Survey National Grid – OSGB36) by sampling from a given set of trial priors and using the prediction model (Equation (10)) with the sampled regression coefficients. The simulated temperature fields were assessed manually to judge their plausibility. If their appearance was sensible then these priors were used. To avoid over precision in the priors we inflated their variance slightly to allow more flexibility in the model as they can now be quickly updated by the data. As such the interpolation model appeared to be insensitive to these initial prior values. Because the UKV forecast (Section 4.4.1) is time-varying a selection of timesteps were used to help set an appropriate initial prior.

Predictors that were tested, but not selected, included estimates of the soil type around a location, how inland a site is, the land/sea ratio in an upwind direction, and how sheltered or exposed a site is with respect to the surrounding topography. This last measure was tested in order to try differentiating stations within valley bottoms from those on top of exposed hills, which can experience remarkably different temperatures under certain synoptic situations such as those conducive to cold pooling. Current numerical forecast models struggle to resolve these effects of local topography (Vosper, et al., 2014). Sheltering effects at the micro-scale will also influence air temperatures; however accurately quantifying this effect for use as a predictor is very difficult. Unfortunately none of these predictors provided significant reductions to model error, while several were computationally expensive to generate, and in the interests of simplicity, all were omitted from the model. The final model only uses the temperature field from the short range forecast model, but other UKV fields were also tested as predictors. These included the forecast soil moisture and cloud cover, but again at most they provided marginal improvements in model accuracy. This is unsurprising as the UKV's temperature forecast should have already accounted for the effect of such influential variables.

4.4.1. Met Office short range forecast

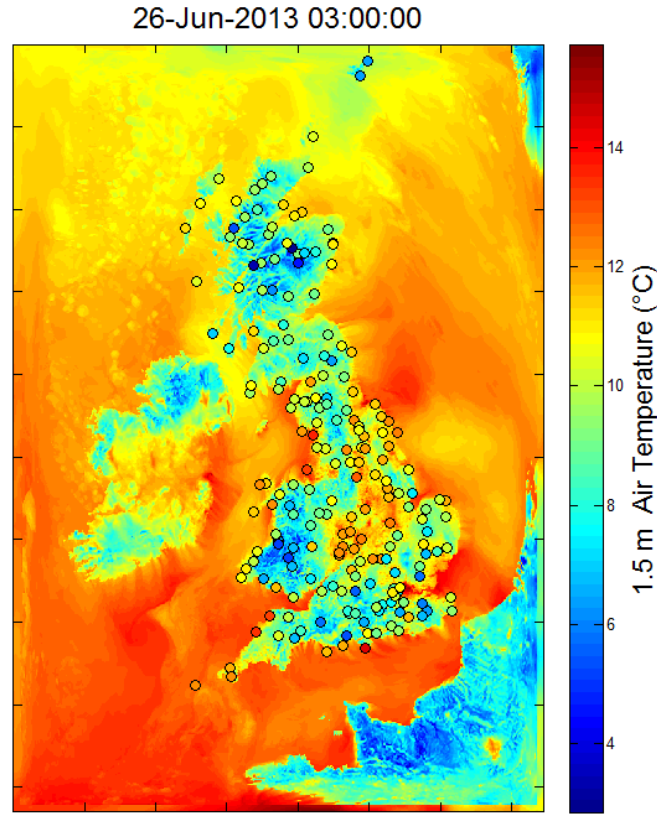


Figure 4.9. Met Office UKV T+3 forecast of 1.5 m air temperature field for 26th June 2013 (summer period) at 03:00. Overlaid with MMS air temperature observations (circles) at the equivalent timestep.

The Met Office’s short range forecast from its UKV model is a powerful predictor in our interpolation model. The UKV configuration is ~1.5 km resolution, ‘convection-permitting’ model covering the British Isles. It is nested within a ~25 km resolution global model using variable resolutions at the model boundaries (Tang, et al., 2013). It comprises of 70 vertical levels on a terrain-following hybrid-height vertical coordinate system. It runs 8 assimilation cycles every day (every 3 hours) using a 3-dimensional variational (3D-Var) data assimilation scheme (Lorenc, et al., 2000). It is a particularly pertinent model to use, since one very likely application of corrected CWS data would be to feed the observations into the model’s high resolution data assimilation scheme. Here we use the UKV forecast at a lead time of T+3 hours. Shorter lead times of T+1 and T+2 are also available, but T+3 was favoured as it is commonly used in model verification and permits model spin-up.

Bilinear interpolation (Lillesand, et al., 2008) of the nearest 4 cells is used to map the gridded model output to the location of each MMS and CWS station. This was done for the values at each 3-hourly timestep. To save processing time we precomputed which 4 cells were nearest to each CWS in advance, along with the distance from each cell

centre to the CWS location. To account for the difference in temperature resulting from the discrepancy between model height and station height, two approaches were tested. The first was to apply a lapse rate correction based on a simple fixed lapse rate of $-6.4\text{ }^{\circ}\text{C per km}$. The second derived a lapse rate from the UKV model itself, by fitting a simple regression model to the relationship between model cells heights and their forecast temperature. This was done for each station using only cells within a $\sim 300\text{ km}$ radius to provide a localised, but stable, estimate of the model lapse rate. For example, if Scotland was experiencing a different air mass, and thus different average lapse rates to Southern England then this approach would apply more appropriate corrections to model cell temperatures in each region. Figure 4.10 shows that both approaches lead to significant improvement in the UKV's RMSE. There is however very little difference between the two approaches. Sheridan, et al., (2010) implemented a similar model-derived lapse rate correction for downscaling from operational mesoscale (4 km resolution) model output. Although they found significant improvements in the temperature correction under stable conditions (e.g. when air temperature increases with height) they also saw limited improvement to the overall RMSE. Because of the marginal difference between the two approaches the fixed lapse rate correction was favoured for its simplicity and significantly lower computational cost.

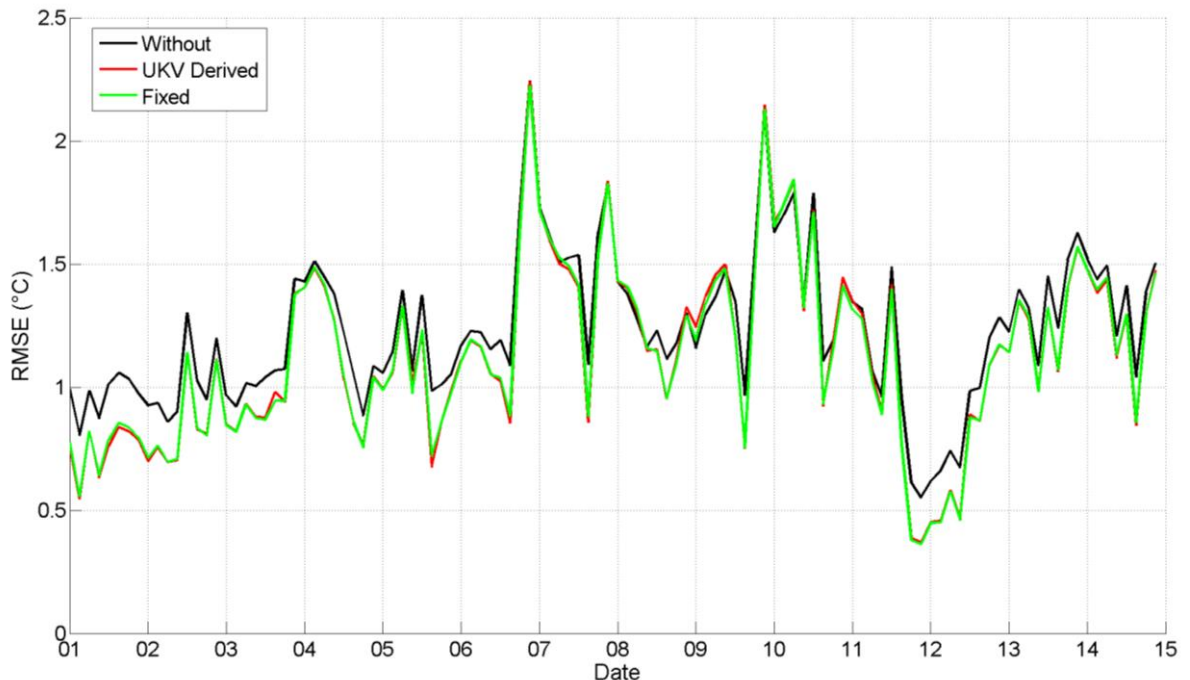


Figure 4.10. UKV RMSE time series during the autumn period (1 - 14 October 2012) for different model cell height to station height temperature corrections; verified against observations from 225 MMS stations over the 112, 3-hourly, timesteps. Vertical grid lines mark midnight.

When including the UKV temperature forecast as a basis function, the roles of the other predictors are altered. For example, the model includes a constant term – without the UKV our expectation of its regression coefficient would be close to the climatological average temperature, but with the UKV we would expect a value of 0. Were it not equal to 0, then this would imply that the UKV systematically over- or under-predicts at every station. It is a similar case with the other predictors. Taking elevation as an example, without the UKV the expectation of its regression coefficient is that it would resemble a common lapse rate, such as $-6.4\text{ }^{\circ}\text{C per km}$. With the UKV included, this basis function now essentially acts to correct any biases in the UKV model which show a correlation with elevation. If there are any such biases then the elevation's regression coefficient would deviate away from 0. The Radial Basis Functions (Section 4.4.5) effectively help to mop up any spatially correlated errors in the UKV model. These other basis functions should also help account for the aforementioned issue of the observations being valid at 10 minutes to the hour while the UKV is valid on the hour. Initially our expectation of the regression coefficient mean for the UKV itself is set as 1 as there is little *a priori* evidence to suggest the UKV has a consistent systematic bias.

The time series plots of RMSE in Section 4.5 illustrate how the RMSE of the UKV varies through each of four case study periods. Figure 4.11 displays the UKVs mean bias at each 3 hour timestep for every day in the summer period. It highlights that the UKV often displays a pattern to its overall bias; here we see signs of a mean cold bias during the day and a warm bias at night. As discussed previously, it is the role of the other basis functions to adjust for these biases.

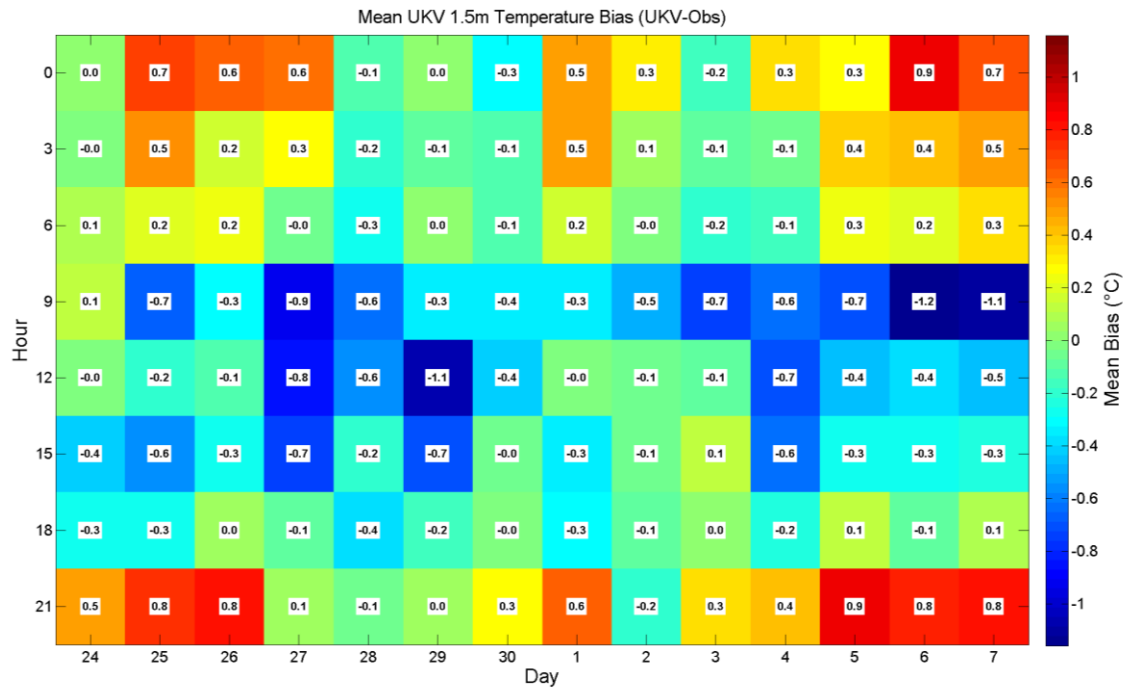


Figure 4.11. UKV bias arranged by 3-hourly timestep (rows) for each day (columns) during the summer period. Verified against observations from 207 MMS stations, of which 20 MMS stations were partially missing data, but were still used when data was available. Values are the mean bias of all the individual station biases at a given time when verified against MMS station observations. Bilinear interpolation is used to map predictions for the 4 nearest grid cells to station locations. A fixed lapse rate correction was used to adjust model predictions to MMS station heights.

4.4.2. Easting, northing and elevation

Each stations' Easting and Northing coordinates, on the Ordnance Survey's National Grid reference system (OSGB36), were used to capture overall north-south and east-west temperature gradients that the UKV may have poorly resolved.

The elevation of MMS sites comes from the Met Office's metadata. They are therefore reliable as supported by Figure 4.12 which shows a strong agreement when verified against Digital Elevation Model (DEM) with a ~250 m horizontal resolution. This DEM is from the GMTED2010 dataset produced by the U.S. Geological Survey (USGS) in collaboration with the National Geospatial-Intelligence Agency (Danielson & Gesch, 2011) who specify a RMSE range between 26 – 30 m. This DEM was re-projected, using bilinear interpolation in esri's ArcGIS software, to a 1km grid in order to simulate a temperature field using sampled elevation regression coefficient values from a trial prior distribution (as introduced earlier in Section 4.4).

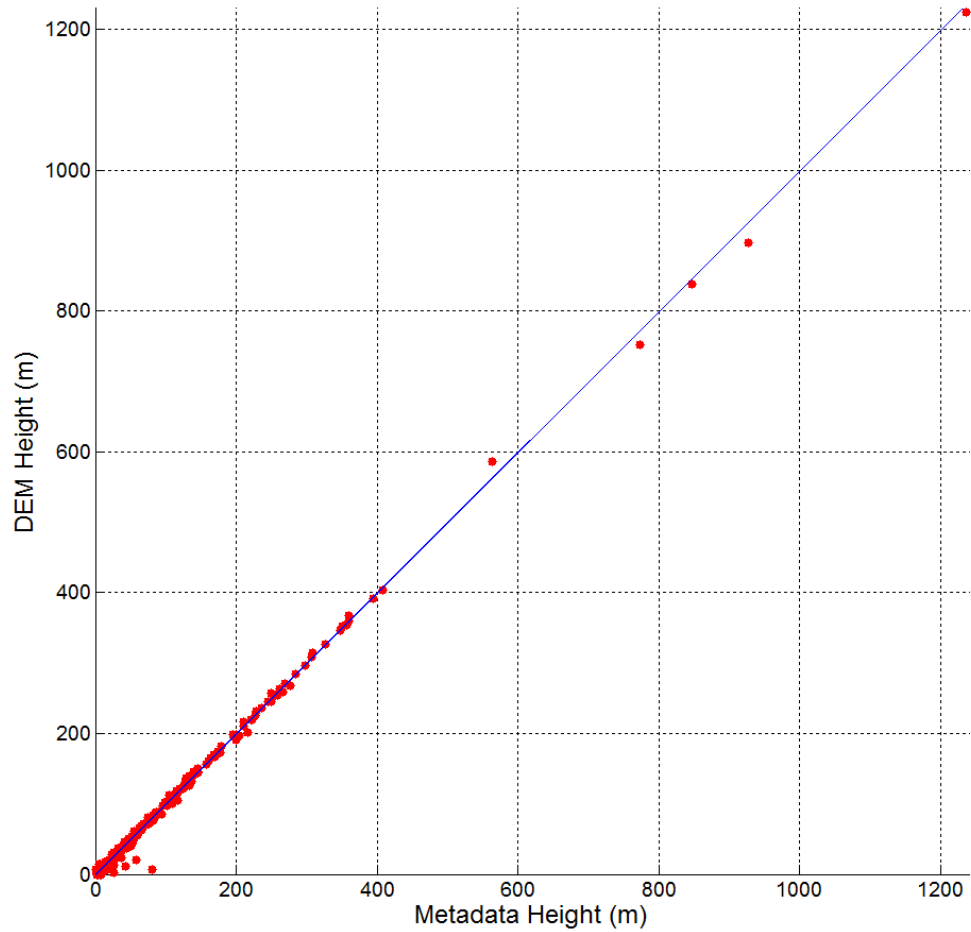


Figure 4.12. Comparison of the MMS station elevations as listed within Met Office metadata with the GMTED2010 DEM derived elevation. The raster DEM was sampled at the station coordinates using bilinear interpolation.

4.4.3. Coastality

To represent coastal impacts on temperature the proportion of sea within a 25 km radius around each station was used. Although coastal effects such as sea-breezes can penetrate as far inland as 85 km (Simpson, et al., 1977), here we use 25 km as it produced the lowest RMSE during cross-validation. Figure 4.13a displays the coastality estimates over the British Isles.

An approach was also tested to estimate the proportion of sea only in an upwind direction, i.e. the direction from which the sensed air parcel has travelled. Distances ranging from 10 km through to 100 km were tested. Unlike the simple 25 km circular radius this upwind approach is reliant on wind direction/speed estimates at every station and every timestep. This is required to either calculate the upwind proportion in real-time, or to assign the prevailing wind conditions to a designated class for which the proportion has already been calculated. Because of these additional computational costs and only a marginal improvement in error this approach was not

included in the model. Such an approach may also be over-simplistic because at times the sensed air mass is a combination of air packets from several directions, not just one. For example, when onshore sea-breezes over a peninsular cause air masses to converge inland (Reed, 2011).

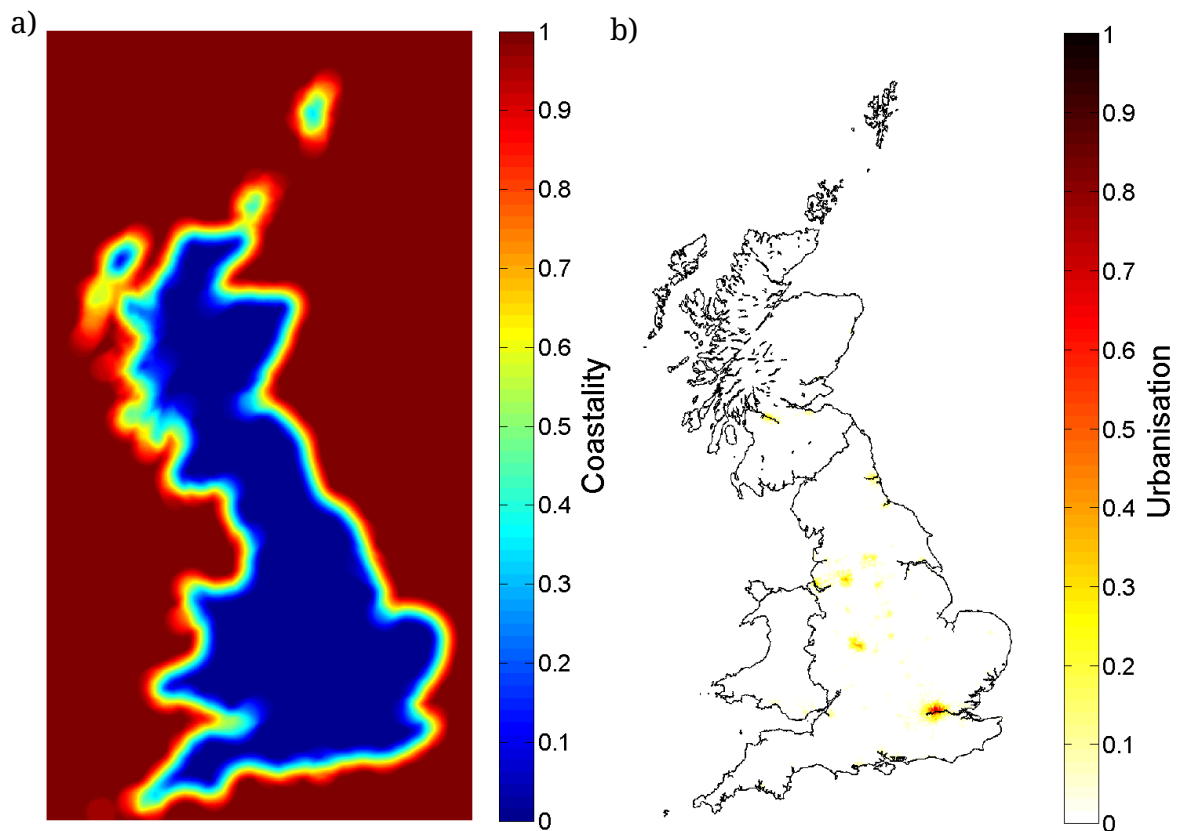


Figure 4.13. a) Coastality and b) Urbanisation estimates across the British National Grid study domain (700×1300 1 km cells).

4.4.4. Urbanisation

Many studies have observed the significant effect of urbanisation on temperature in UK cities (Jones & Lister (2009); Smith, et al., (2011); Tomlinson, et al., (2012)). Even though the UKV model has a resolution of ~ 1.5 km, and should resolve some of the effects of urban areas (Best, 2005), it may still display biases that are correlated with urbanisation. This basis function acts to correct such biases.

The degree of urbanisation around each station was derived from the LCM2007 25m Raster land cover map (Morton, et al., 2011). This very high resolution map differentiates between *urban* and *suburban* areas. *Urban* areas include town and city centres, where there is typically little vegetation, as well as dock sides, car parks, and industrial estates. *Suburban* areas comprise a mix of urban and vegetation signatures. The map was reconfigured so that *urban* cells have a weighting of 1, *suburban* 0.5 and

every other land cover type 0. *Urban* is weighted more heavily than *suburban* as *urban* areas typically displaying more pronounced urban heat island effects (Stewart, et al., 2014). Several relative weightings of *urban* to *suburban* were tested, with minimal differences to the overall model error, thus 1 and 0.5 respectively were selected. The mean of the cells within a given radius around each station was then calculated. This was done for two separate radii, 1 km and 10 km; the means were then multiplied. The theory behind using these two separate sizes is that it now not only considers the degree of urbanisation locally, but also the size of the town/city the station is sited within. Figure 4.13b shows the calculated degree of urbanisation across the whole domain. Note that because of the larger radius locations within large cities receive a much stronger degree of urbanisation than small towns and villages. In reality however there was little difference in RMSE when the interpolation model was cross-validated using this double-radius approach over a single localised estimate.

4.4.5. Radial basis functions

RBFs help to learn and predict any localised temperature fluctuations or spatially correlated biases within the other basis functions. MATLAB's K-means clustering function was used to space the centres of Gaussian RBFs evenly over only the land, i.e. by only using the terrestrial easting and northing coordinates. Were a regular grid used instead then many centres would have been placed over the sea with few nearby stations to learn from. 20 RBFs were used as it produces a sensible spread of centres across the country (Figure 4.14) and led to a marginally lower RMSE in comparison to different numbers of RBFs. The squared mean minimum distance between centres was used as the variance for all the RBFs. It was multiplied by 0.7 first so that the RBFs would be localised. As with the number of RBFs, various values for this multiplication factor were tested, with 0.7 selected as it led to a marginally lower RMSE.

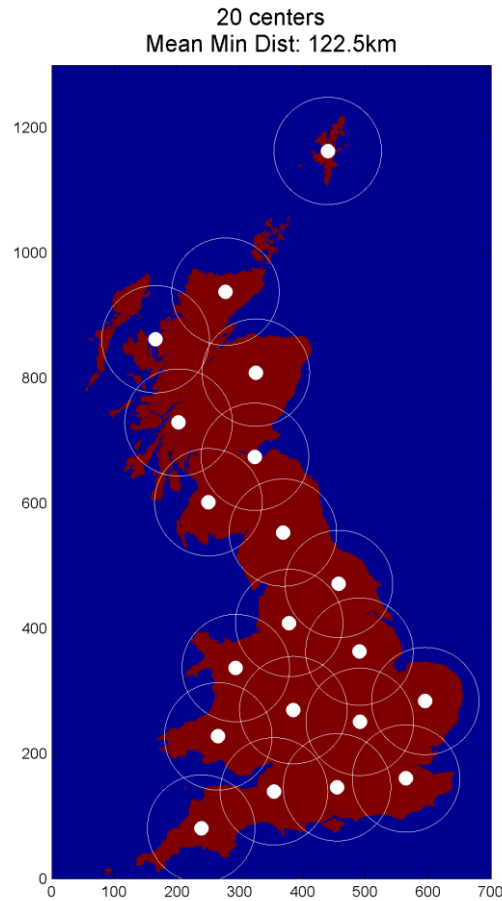


Figure 4.14. Distribution of RBF centres having use K-means to locate them over land areas only. Rings around each centre represent 1 standard deviation.

4.5. Model performance

Before using this interpolation model to estimate the temperature at CWS locations, it is vital to assess its performance. The model's temperature estimates should fall within a satisfactory level of error, be unbiased on average, and crucially should validate well probabilistically. As each temperature estimate has an associated uncertainty value that propagates through to the bias correction model it is important that we are not systematically over- or underconfident with these estimates.

10-fold cross-validation (Bishop, 2007) was used to verify the model. The MMS stations were partitioned into 10 groups and each group in turn was withheld, the model was then trained using the other 90% before predicting at, and validating against, the observations from the held-out 10%. The group allocations were assigned randomly, but could be fixed when testing the model's sensitivity to other factors. The allocation of stations to particular folds had little impact on the overall accuracy of the model, although it is important to ensure that each RBF is informed by a significant number

of stations. 10-fold was favoured over a 'leave-one-out' approach to reduce computational time.

Overall the interpolation model performed satisfactorily. Figure 4.15 is a 1:1 plot displaying the model predictions against the MMS observations for every station, timestep, and case study period, combined. The closer the points fall to the 1:1 line the more accurate the model is. For all the four case study periods combined the mean bias of the model is 0.00 °C, with a residual variance of 0.68 °C, and RMSE of 0.82 °C (Table 3). As the mean bias is so close to 0 °C the standard deviation of residuals is virtually identical to the RMSE and is therefore omitted from Table 3. It is interesting that the largest errors occur at lower temperatures, evident by the greater spread around the 1:1 line. This may be a result of cold stable conditions when source areas are more localised making accurate interpolation more difficult than under well-mixed conditions. The histogram of the residuals (Figure 4.16) confirms that overall the model is unbiased with a distribution of residuals virtually symmetrical about zero.

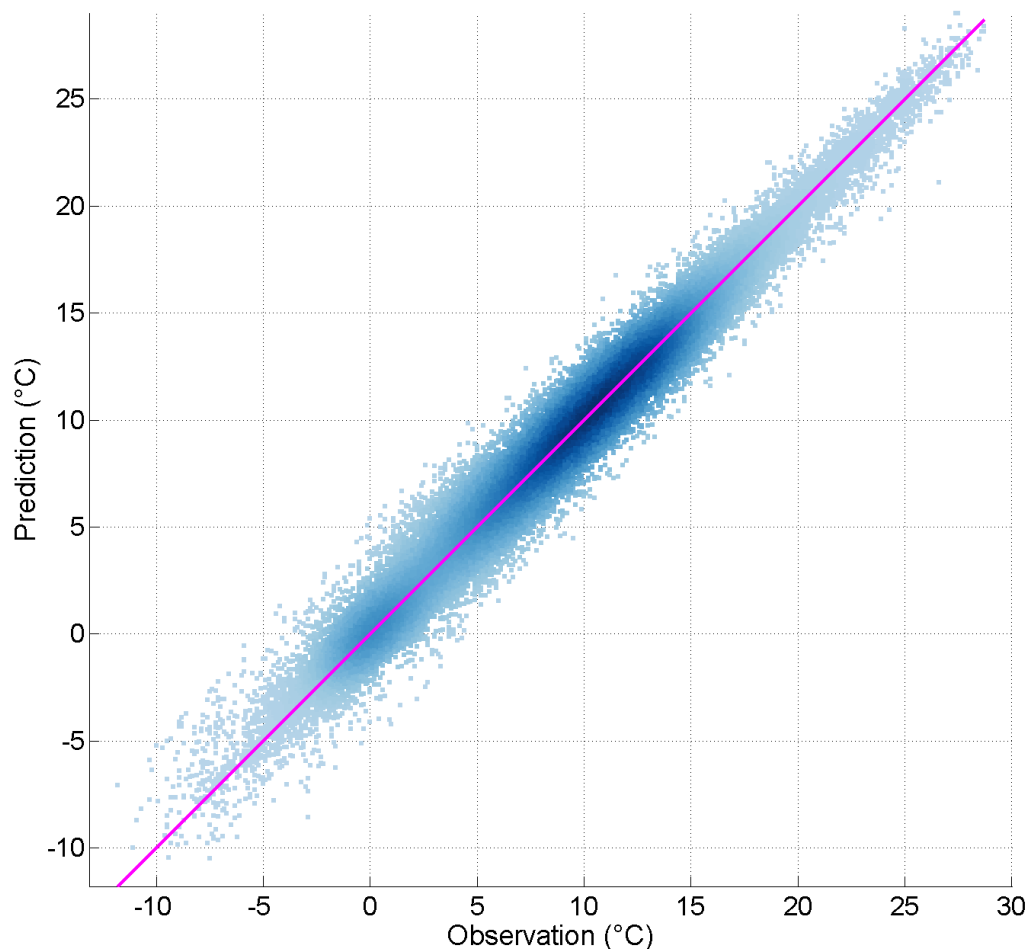


Figure 4.15. 1:1 plot of interpolation model temperature predictions against MMS observations using 10-fold cross-validation. Data shown is for all four 2 week periods; run separately, but plotted together. Magenta line is a 1:1 line. The darker the colour the higher the density of points.

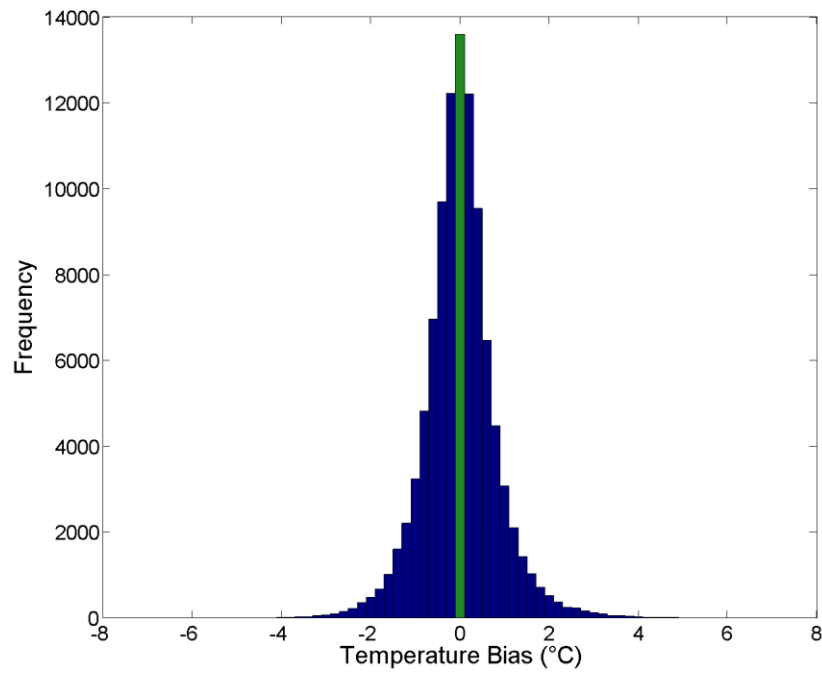


Figure 4.16. Histogram of residuals when the interpolation model was verified against MMS observations using 10-fold cross-validation. The residuals for all four periods combined are shown.

With an RMSE of 0.82 °C the model's accuracy is acceptable, but only if it validates well probabilistically. For this to be the case, an accurate temperature estimate should also have a corresponding uncertainty estimate which is low, and conversely when the model makes poor estimates its uncertainty estimate should be high. This can only be verified statistically, i.e. using multiple points to check that on average this behaviour is evident. Figure 4.17 through to Figure 4.20 help illustrate the model's probabilistic performance. The distribution of z-scores and the shape of the rank histogram imply we are not systematically over- or underconfident with our predictions; at least not significantly. The points in the coverage plot (Figure 4.19) and reliability diagram (Figure 4.20) lie close to the red 1:1 lines confirming that the residuals from the interpolation model closely follow a Gaussian distribution with the defined variance. Technically the residuals follow a Student-t distribution; however as the sample size is large enough they can be closely approximated by a Gaussian distribution as shown here. A Gaussian distribution is favoured for ease of use.

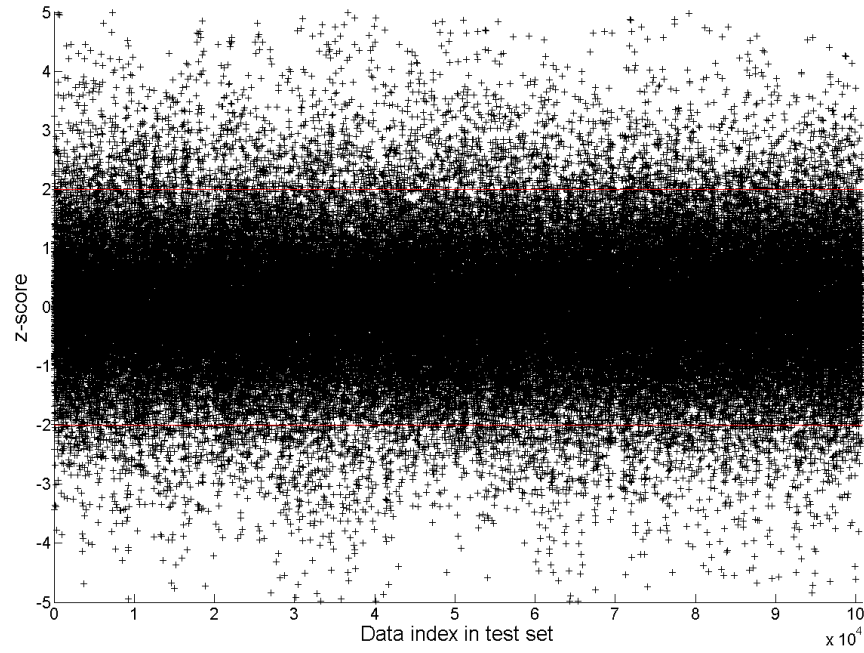


Figure 4.17. Plot of z-scores for all four periods combined. Ideally ~95% of the z-scores should fall within the two red lines at ± 2 . Here 93.5% fall within this range. The z-score, $z = \frac{x - \mu}{\sigma}$, where x is the MMS observation, μ is the interpolation model's mean prediction, and σ is the estimated standard deviation of the prediction (i.e. the estimated uncertainty).

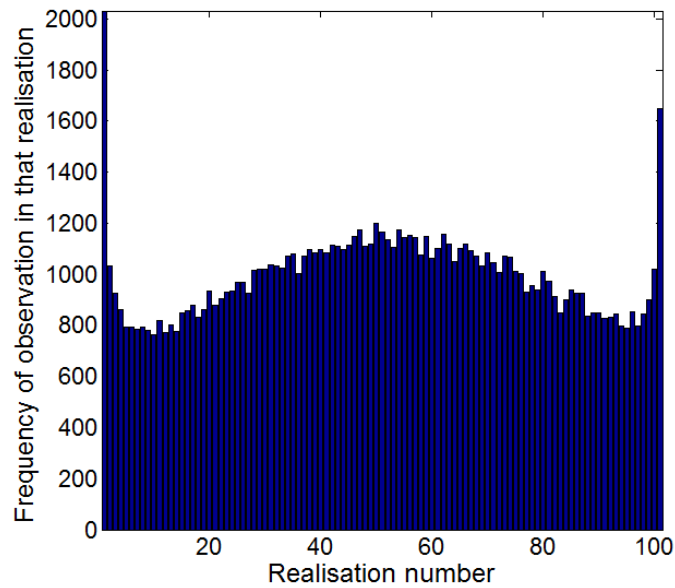


Figure 4.18. Rank histogram for all four periods combined. A flat, uniform, appearance implies a good probabilistic model (Hamill, 2001).

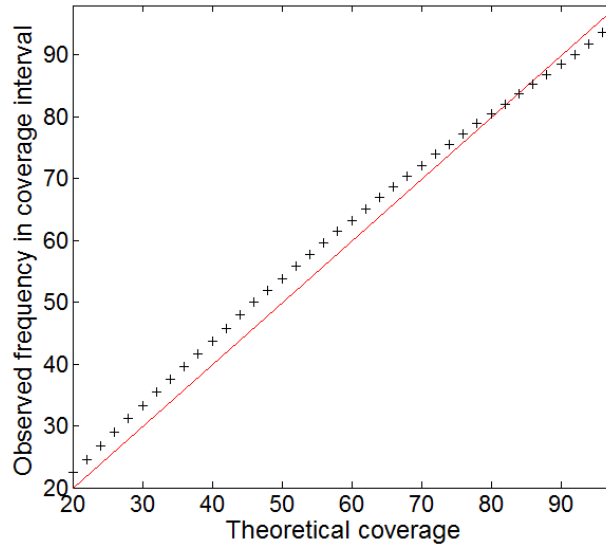


Figure 4.19. Coverage plot for all four periods combined. It plots the theoretical centred confidence interval against the observed frequency. When the points fall close to the red line the model has validated well probabilistically.

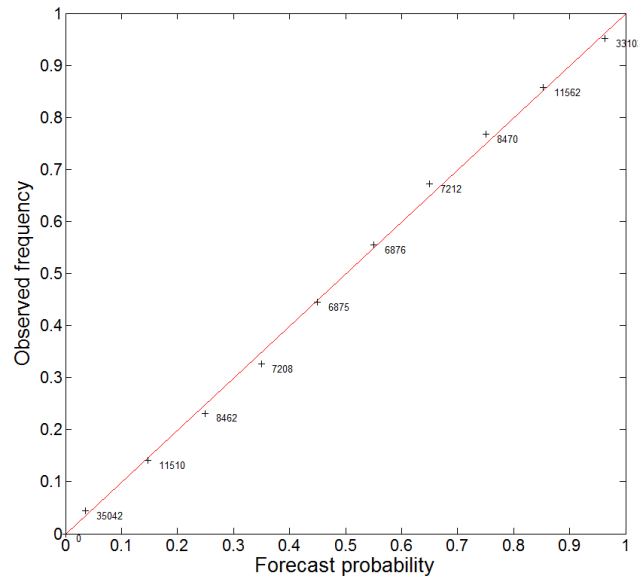


Figure 4.20. Reliability diagram for all four periods combined. This diagram compares forecast probabilities with actual observed frequencies (Bröcker & Smith, 2007), computed by splitting the range of observations into 10 classes. The number attached to each point denotes the number of observations in each class.

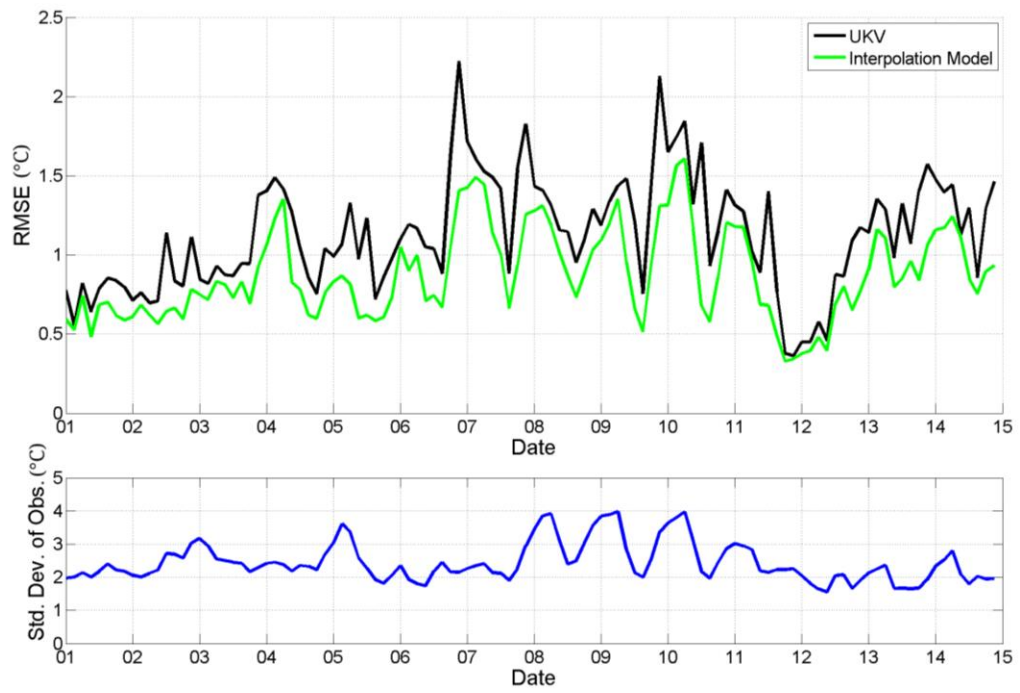
If we look at the statistics of each period separately (Table 3) it is clear that overall the model remains unbiased within every period, but has subtle differences in the spread and magnitude of the error. In Figure 4.15 we saw that the model displayed some large errors when it tried to predict low temperature observations experienced during the winter period. Despite this, the winter period is actually the better predicted period with the lowest residual variance, RMSE and MAE.

Table 3. Error statistics for the interpolation model during each of the four 2 week case study periods. Verified against MMS observations using 10-fold cross-validation.

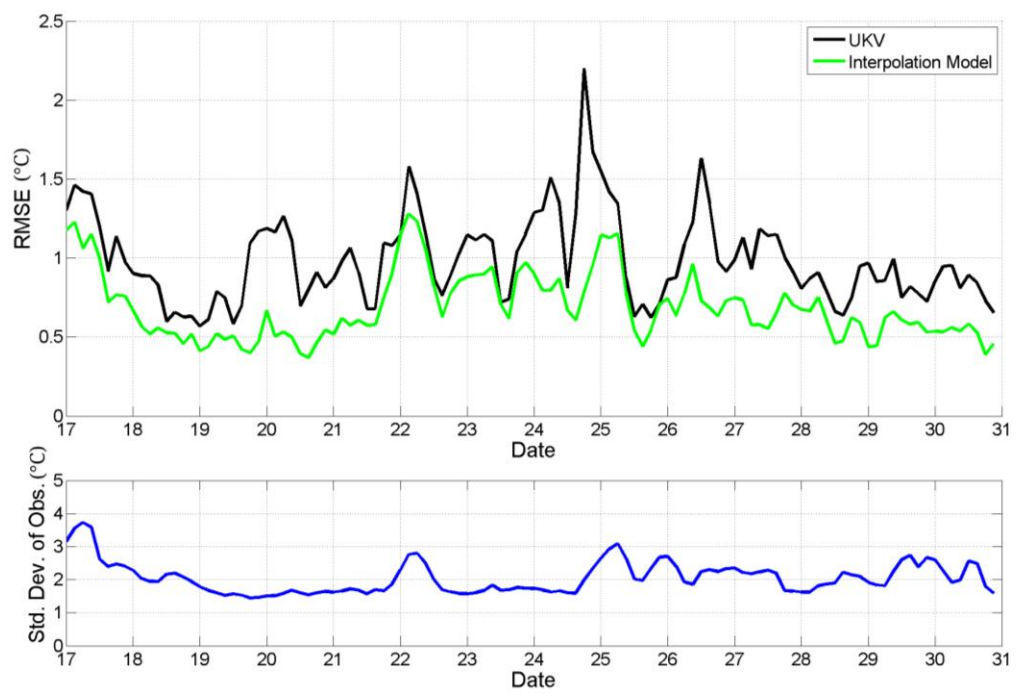
	Mean Bias (°C)	Residual Variance	RMSE (°C)	MAE (°C)
All Periods	0.00	0.68	0.82	0.59
Autumn	0.01	0.84	0.92	0.65
Winter	0.00	0.52	0.72	0.50
Spring	0.00	0.70	0.84	0.61
Summer	0.00	0.64	0.80	0.69

Figure 4.21 plots the RMSE through time for each period, both for the UKV model on its own and for the full interpolation model in which the UKV is just one of many predictors. It is clear that the inclusion of the other basis functions leads to a significant improvement over using just the UKV output alone. The standard deviation of the observations is also plotted. From this we can see whether poorly predicted situations are simply a result of high variation in temperature observations at the time. There is certainly evidence that large RMSEs occur when the standard deviation of the observations is also high - however, this is not the full story as there are several examples of high RMSEs when the standard deviation is low; for example, around midnight on the 7th October 2012 (autumn period) and around 03:00 on the 6th July 2013 (summer period). There are signs of a diurnal pattern to the model error, but the time of day when the peak error occurs varies between, and even within, periods. As detailed earlier in Section 2.3.5 the source area measured by a thermometer varies with the wind and atmospheric stability, which vary through time. When the source area is small (e.g. during unstable conditions) the interpolation model is likely to struggle. This effect may explain the large RMSEs even when the standard deviation is low, although further investigation is needed to prove this possible correlation with stability.

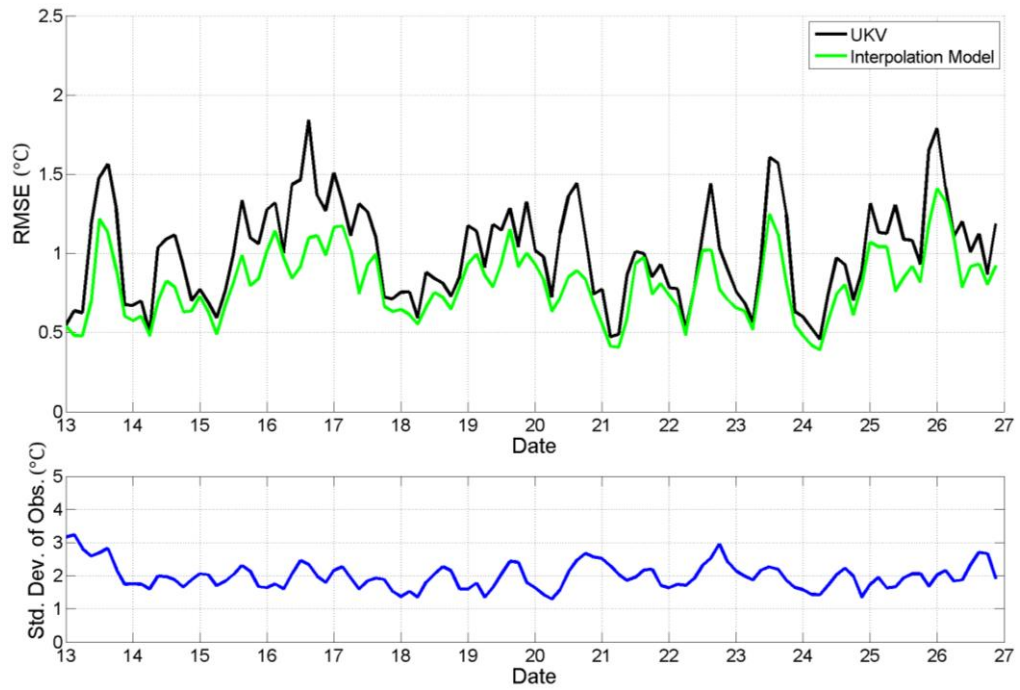
a) Autumn



b) Winter



c) Spring



d) Summer

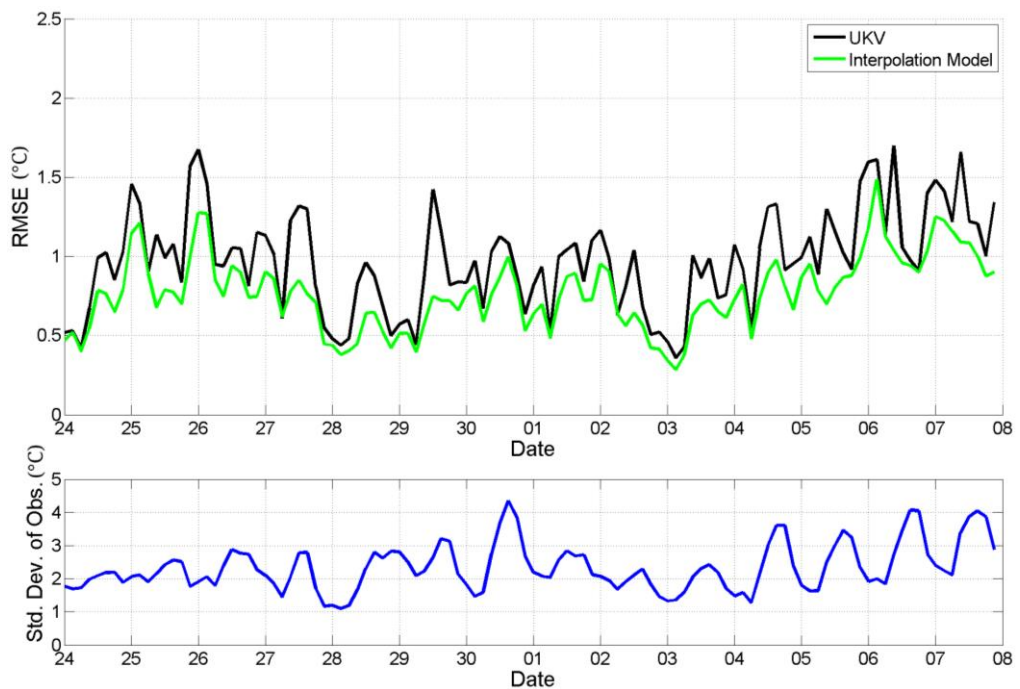


Figure 4.21. Time series of root mean squared error (RMSE) of the UKV model and the Interpolation model over each of the four 2-week case study periods. Both models were verified against the same set of 3-hourly air temperature observations from the ~220 MMS stations. Below each plot is a time series of the standard deviation of the observations. Vertical grid lines represent midnight at the start of each date.

The model's accuracy varies not only through time, but also spatially. Figure 4.22 shows that in each period certain stations are better predicted on average than others.

Stations located in central England tend to display the lowest average errors with those in coastal and mountainous regions proving most difficult to predict at.

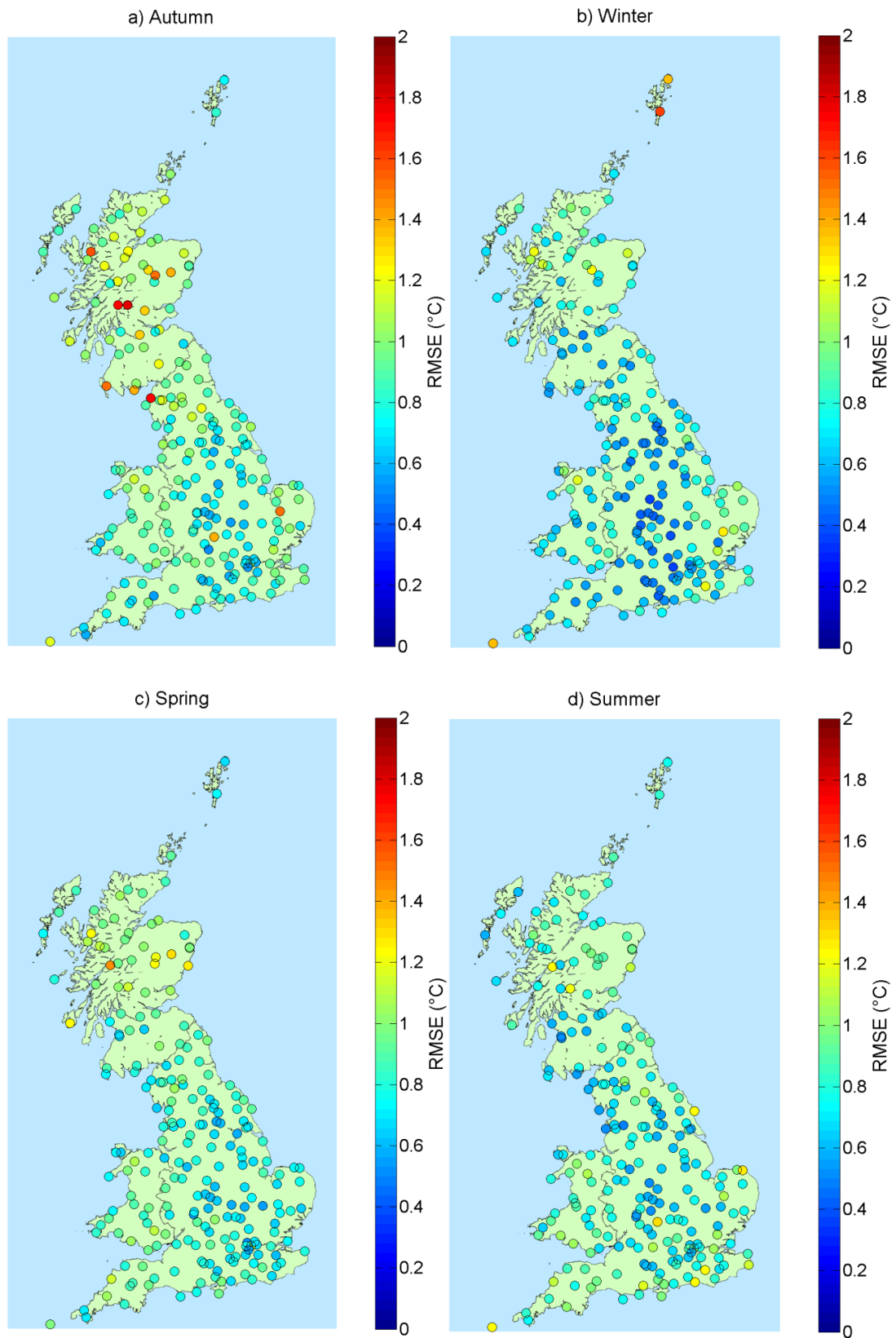


Figure 4.22. Interpolation model RMSE of predictions made at the location of each MMS station. RMSE is calculated from the residuals of every timestep within the given period.

When the model errors are averaged over time, as in Figure 4.22, it is easy to hide significant errors that may only occur at a single timestep. Therefore, Figure 4.23 displays model errors at a single timestep, namely the 10th October 2012 06:00 during the autumn period. For comparison the observations, the UKV forecast, and the mean interpolation model prediction are also shown. This timestep was the worst predicted out of all four 2 week periods with an RMSE over 1.5 °C and errors at individual stations in excess of ± 3 °C. It is interesting that there is little evidence of a spatial structure to errors, i.e. significantly over-predicted stations are often not far from those that have been under-predicted. During this particular timestep Great Britain was still in full darkness and was covered by patches of low-level cloud. There was a large variation in temperature (std. dev. of 4 °C) ranging from over 10 °C to below freezing with night frosts likely in places. Clearly this proved to be a difficult situation for the UKV model, and subsequently for our interpolation model.

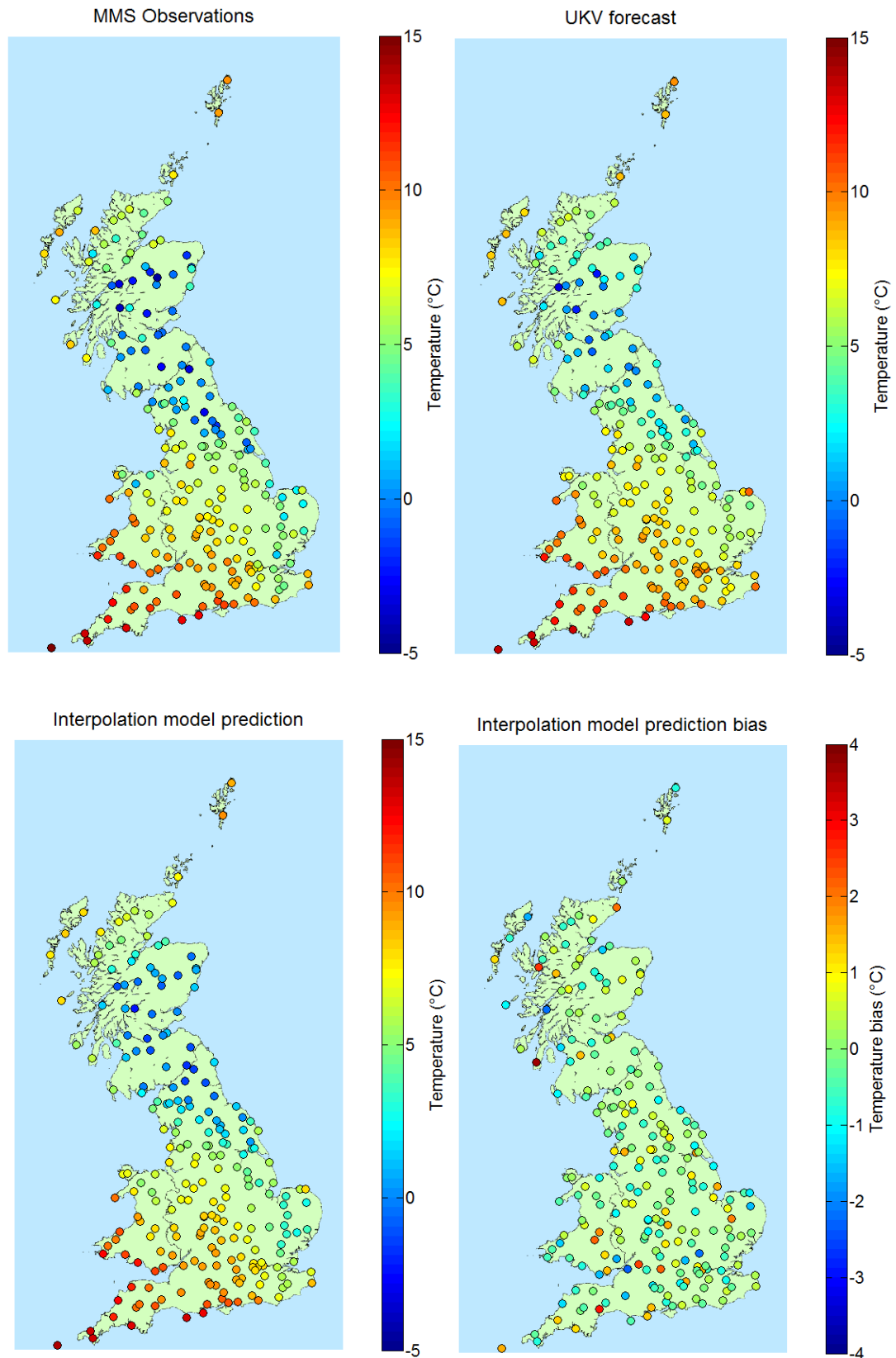


Figure 4.23. Spatial plots for the case study timestep of 10th October 2012 06:00 (autumn period). The figure plots the MMS observations, the UKV model's predictions (height correction applied), the prediction from the interpolation model and its bias when verified against the MMS observations using 10-fold cross-validation. Note that the bottom right plot shows

temperature bias rather than absolute temperature, and thus the colour scale has changed to accommodate for this.

Figure 4.24 illustrates the model error statistics when verified against each individual station. For most stations the median bias is close to zero with an interquartile range close to 1 °C. There are however stations for which the interpolation model has greater difficulty predicting; for example stations at high elevations. Also note the poorly predicted station with a relatively low elevation, but with its full interquartile range below 0 °C with a median of 0.77 °C. This station is called Cromer and lies just 100 m away from the north East Anglian coast. Our interpolation model consistently under-predicted the temperature at this location, further illustrating that temperature is hard to predict at coastal stations. These biases imply either error in the model, or that the given station is not representative of the scale at which the model resolves.

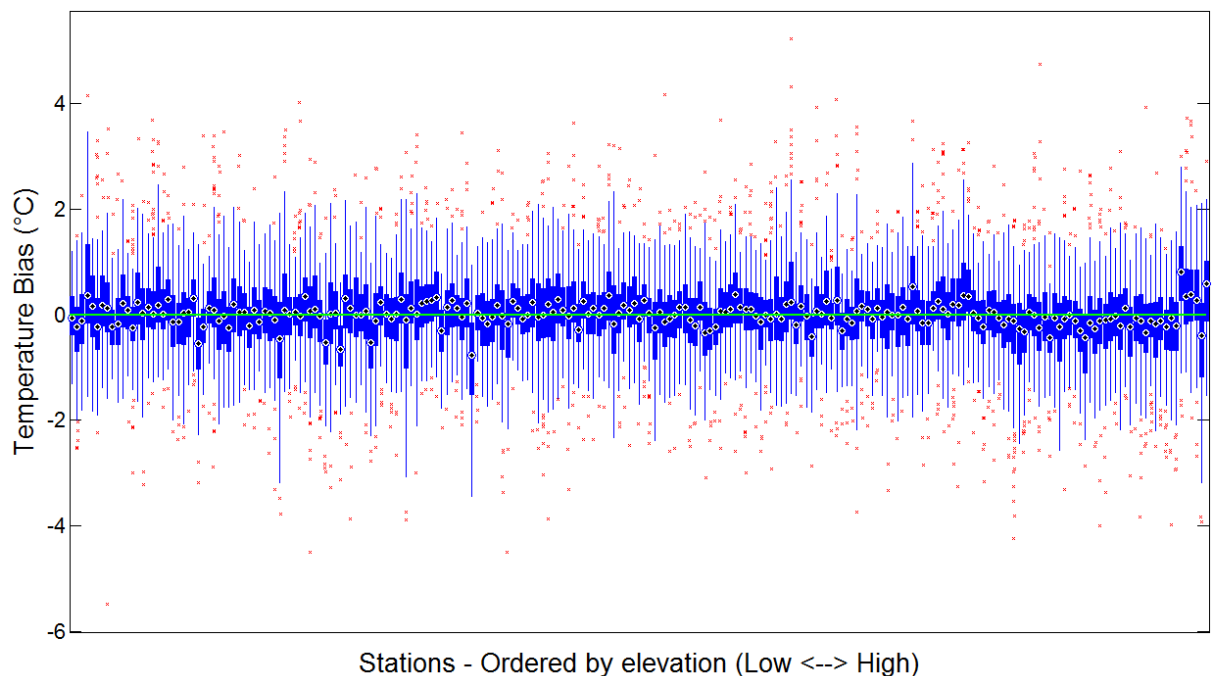


Figure 4.24. Interpolation model error statistics when verified against each individual MMS stations' observations using cross-validation. Results here are for the summer period only. The order of the stations is based upon their elevation. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

As was detailed in Section 4.3.2 the regression coefficients are allowed to evolve and vary through time, informed in part by what was learnt at previous timesteps, but updated as new information arrives. An alternative approach would be to learn the regression coefficients using only the new training data, ignoring what was learnt previously. Figure 4.25 shows, that in terms of the model's RMSE, there is virtually no

difference between these two approaches. Because a lot of new training data arrives at each timestep and because the forgetting rate parameter, γ_β , is set relatively short (24 hours) it is likely that the prior information brought forward has very little influence on the update of the regression coefficients. Were γ_β increased then the learnt regression coefficients from preceding timesteps would have greater influence, meaning that the regression coefficients would evolve more slowly. However, dramatically increasing γ_β tends to increase the model's error slightly. This makes sense as we would expect that the regression coefficients would need to evolve quickly to account for often swift changes in the prevailing weather conditions. For example, changes due to a frontal passage or a switch between onshore and offshore breezes can easily occur over the space of a day. As such, setting γ_β no larger than 24 hours is appropriate. Despite the marginal difference, the propagation approach was still favoured as it provides an elegant solution to handle any timesteps with limited amounts of new training data (e.g. due to missing data) where the learnt regression coefficient from preceding timesteps should be more influential, ensuring the stability of the model. This feature would be particularly useful if we were to run the clustered approach, as detailed in Section 4.3.3, as at times a cluster may become very small, informed by very few stations.

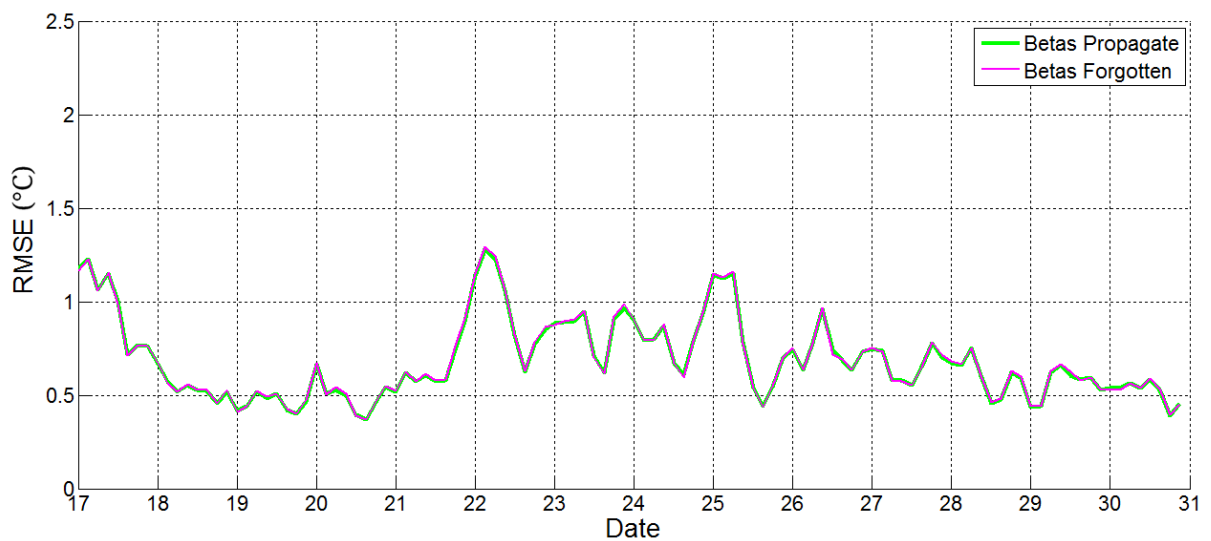


Figure 4.25. Time series of the interpolation model cross-validation RMSE during the winter period, both when the regression coefficients are allowed to propagate through time, and when they are forgotten after each timestep and then re-learnt from scratch at the next timestep.

Figure 4.26 demonstrates this evolution of the regression coefficients using the mean term of the UKV's regression coefficient as an example. Note how it is allowed to fluctuate relatively quickly through time. What is reassuring is that the values are

consistent between cross-validation folds which have been trained using slightly different combinations of stations. As it is unlikely that the UKV would display any significant long-term systematic biases, it may seem peculiar that the regression coefficient mean values are not closer to a value of 1. Were the UKV the only predictor then this would be the case, however with the other basis functions included the UKV regression coefficient must adjust in balance with the other regression coefficient values in order to produce the tightest model fit to the training data.

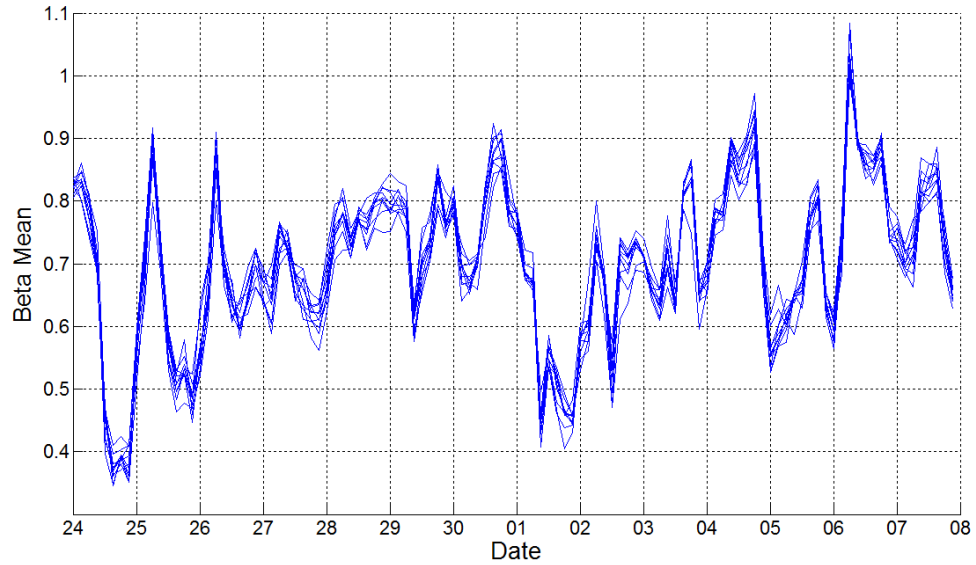


Figure 4.26. Time series of the regression coefficient mean for the UKV basis function over the summer period. Each line represents a different fold within the 10-fold cross-validation.

As a side topic, we theorised that when the UKV model and our interpolation model (without the UKV included) disagreed on the temperature at a given location/time, our interpolation model (with the UKV included) would therefore be more likely to produce a poor estimate. If this were the case, then we would be able to identify in advance situations when we would be likely to make a poor prediction. In order to quantify the discrepancy between the UKV and interpolation model predictions, the Hellinger distance was used in order to also account for the uncertainty estimates of each prediction. The Hellinger distance quantifies the similarity between two probability distributions (Gibbs & Su, 2002); in this case, two Gaussian distributions. For the UKV this uncertainty estimate was set using the residual variance at a given timestep when verified against MMS observations. Unfortunately, as Figure 4.27 shows, there is no obvious relationship between the Hellinger distance and the interpolation model error. Therefore, consistency between interpolation model and UKV predictions does not necessarily imply a good prediction (conversely inconsistent predictions do not imply a bad prediction). This suggests that the noise is truly random.

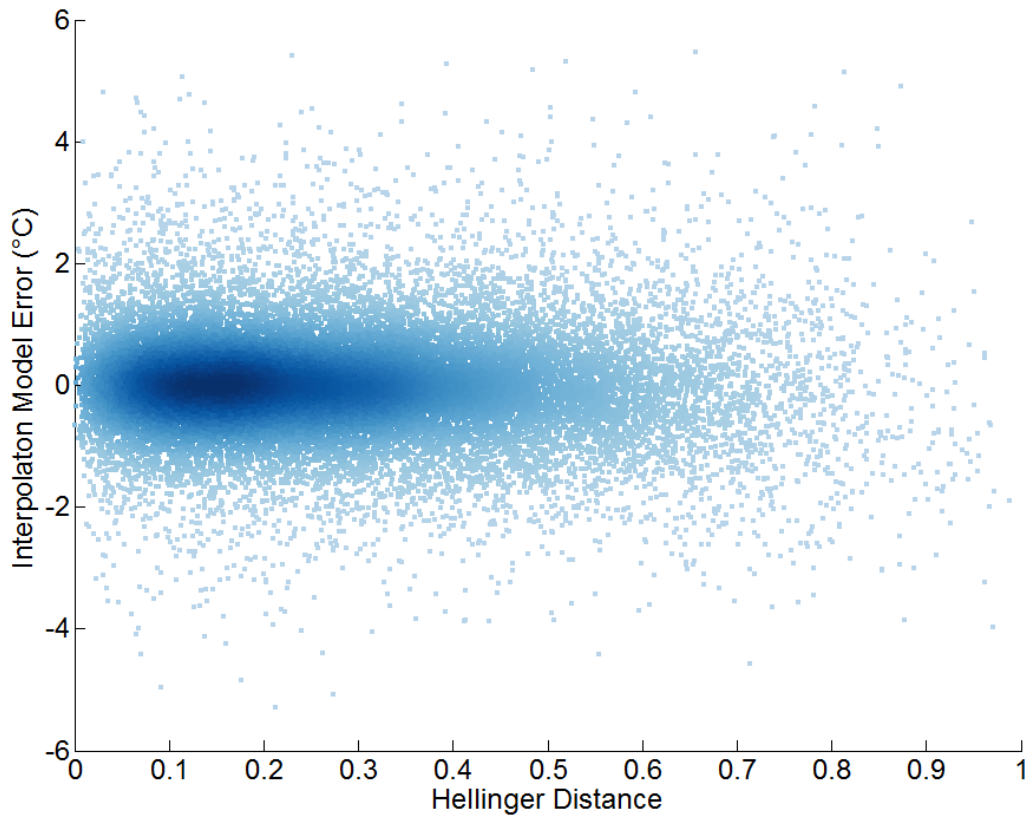


Figure 4.27. Relationship between the UKV and interpolation model discrepancy (as quantified by the Hellinger distance) and the interpolation model error over the autumn period. The closer the Hellinger distance is to 0 the greater the agreement between the two models.

4.5.1. Predictive power

Given the range of different basis functions used within the interpolation model, it is useful to quantify which has the greatest impact on the model's performance. One approach to assess each predictor's impact is to leave that particular predictor out and see what influence this has on the model error. Although more robust metrics such as ANOVA (Analysis of variance) should be considered in the future, this approach still proves to be informative. In Figure 4.28 we show the impact on the RMSE time series over the summer period. It is clear that removing the UKV causes the greatest increase in the RMSE, implying that it provides the greatest predictive power. For this particular period its impact was often greatest during the day. When removed, the RMSE was actually of a similar order of magnitude to that achieved when solely using the UKV forecast.

The RBFs, elevation and the constant term also provide significant contributions. The RBFs are the only basis functions that at times cause a significant increase in the model error by their inclusion, however their overall value in improving predictions far outweighs this occasional disadvantage. It is surprising that the other basis

functions provide so little benefit to the model. It appears that the UKV resolves these factors well enough already, at least in this relatively simple linear model.

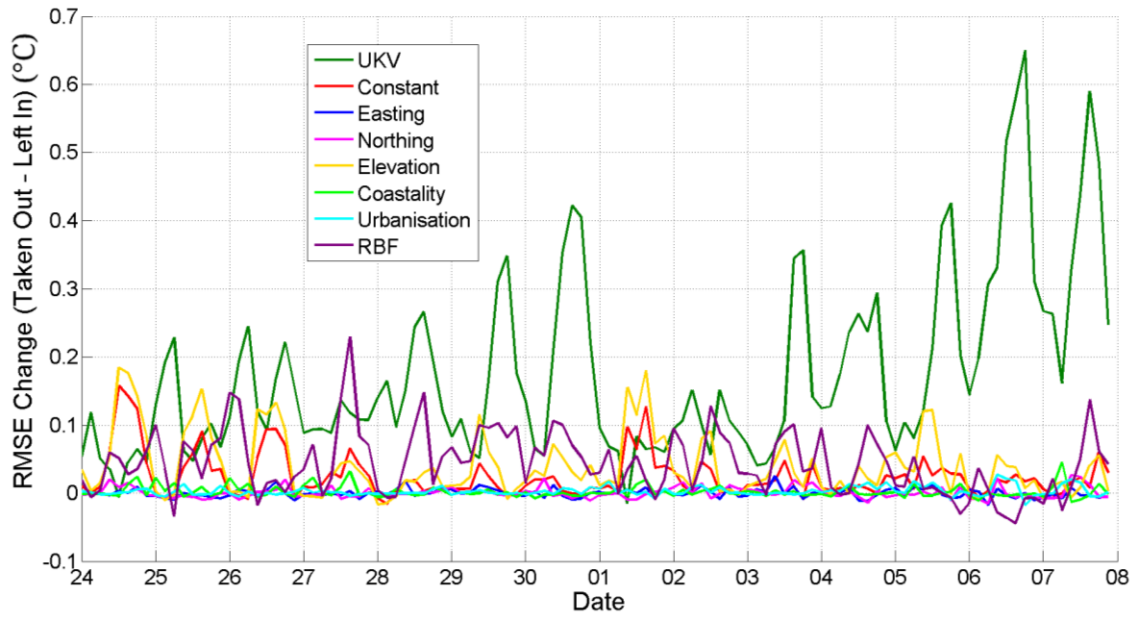


Figure 4.28. Each time series represents the change in RMSE when the given predictor is removed from the interpolation model; in this case for the summer period. A positive RMSE change implies the model accuracy decreases when the given predictor is left out.

4.6. Summary

In summary, we have successfully developed and tested a spatial interpolation model capable of interpolating professional temperature observations across the British domain. As Figure 1.2 illustrates, it can now be used to provide temperature estimates at the CWS locations that are fed into the bias correction model (Section 5.6). Cross-validation verified that the model performs with an acceptable degree of error under a variety of different synoptic conditions throughout the year. Crucially the temperature estimates are unbiased (on average) and have reliable associated uncertainties. Given that these uncertainties are propagated through to the bias correction model (Section 5.6) it was vital that the model validated well probabilistically. The verification illustrated the importance of the UKV as a predictor. Given that the UKV only produces forecasts at hourly lead times and on 3-hour assimilation cycles, further work is required if the aim is to bias-correct CWS observations at sub-hourly resolutions. Further work is also required to adapt the model to handle other variables such as relative humidity and precipitation. We expect that the current approach could interpolate relative humidity relatively successfully if it was handled in the form of dew-point temperature so that errors resemble the Gaussian distribution assumed. A different approach would probably be

required to interpolate precipitation, for example using a radar-guided approach (DeGaetano & Wilks, 2009).

This interpolation model is now be used to produce an independent estimate of the temperature at real CWS locations against which their bias can be learnt. From now on these interpolated MMS observations, interpolated to the CWS locations, are referred to as IMMS for brevity.

5. Modelling citizen station bias

This chapter presents an approach for quantifying the biases present within CWS temperature observations. The chapter begins by describing the real CWS data the model attempts to bias-correct (Section 5.1), before comparing these uncorrected observations against MMS observations interpolated to their locations, IMMS (Section 5.2). The aim of this exploratory analysis is to highlight the magnitude of the differences and identify any obvious bias tendencies evident in operational data. We can check whether the biases apparent within real CWS data agree well with those identified during the intercomparison field study. As discussed below our bias correction model handles calibration biases and radiation-induced bias separately. Estimating the later requires an estimate of the strength of incoming solar radiation at every CWS location. The interpolation model used to produce these estimates is discussed in Section 5.3. As demonstrated in the field study, the magnitude of these radiation-induced biases show a dependency on station design. Section 5.4 discusses how learning the station type can help to anchor the bias correction model's estimates of radiation bias. Even if our bias correction model successfully learnt and removed calibration- and radiation-induced biases, there would still be differences when the corrected CWS data is compared against IMMS values. This is in part due to the natural spatial variation we wish to capture, but may also result from representativity errors. For example, the estimate made by the temperature interpolation model may be representative of a different scale to that which the CWS observes; producing a difference between the two. This challenge of handling representativity is discussed in Section 5.5. In Section 5.6 the bias correction model structure is explained in detail. Then in Section 5.7 we assess the performance of our complete model.

5.1. Input CWS data

The complete bias correction model detailed within this chapter is tested with real CWS data. The data chosen comes from the Met Office's WOW website (wow.metoffice.gov.uk); extracted over the four 2 week case study periods listed in Section 4.2. As the WOW website does not support bulk data downloads a web scraper (as introduced earlier in Section 2.1) was implemented to gather the required data. Through this process we were able to extract around 400-600 stations' worth of data for each period.

As mentioned in the previous chapter, the hourly MMS observations used within the interpolation model are valid at 10 minutes to the hour. To ensure continuity between

the MMS observations and CWS observations only WOW data between 16 minutes and 4 minutes to the hour was used. For stations with several observations in this window the closest to 10 minutes to the hour was selected. As in the intercomparison field study (Section 3) it is unclear whether the CWS temperature observations are point samples or averages. However, as the transmission frequency of all the CWS tested was less than 60 seconds we assume that the observations are valid for the recorded timestamp. We also assume that a CWS's recording rate is more frequent than, or at least equal to, the upload rate to WOW. The data on WOW is prone to missing observations. The bias correction model (Section 5.6) can account for such data gaps, and stations with very high proportions of missing data will be excluded (Section 5.6.2).

It is important to reiterate that the UKV forecasts are instead valid *on* the hour, as are the satellite images used in Section 5.3.2. This timing mismatch may induce slight errors; however our bias correction model should also account for this additional uncertainty and it is more important that the CWS data is valid at the same time as the hourly MMS data, which unlike its minute resolution counterpart has undergone strict Met Office quality control procedures.

As the temperature interpolation model (Chapter 4) relies on UKV data available every 3 hours, we only extract WOW data at the same resolution. Given that WOW data is often uploaded at 5 and 1 minute resolutions, a lot of CWS data is ignored in this study. Making use of all the CWS data is a subject for further work.

In addition to scraping the observations the web scraper also extracts each station's metadata (Section 2.1.4). The key information used includes the station's coordinates, its elevation, site ratings (e.g. for exposure and Urban Climate Zone), as well as the textual *Site Description* and *Additional Information* from which the station type is derived (Section 5.4.1).

5.2. Exploratory analysis of WOW and MMS data

Having developed a reliable temperature interpolation model (Chapter 4) CWS observations can now be compared against an independent estimate of the temperature at their location. This section simply compares these interpolated estimates, IMMS, against our uncorrected input CWS data from WOW. Although this CWS data has not been passed through the full quality control system explained later in the chapter some basic quality control checks (Section 5.6.2) have however been used to remove obvious gross errors. Differences between the two can arise from the combination of many of the following reasons: natural spatial variations, calibration

and/or radiation-bias in the CWS data, error in the interpolation model, and representativity errors. In later sections of this thesis we begin to quantitatively tease apart the influence of each of these factors.

Figure 5.1 shows an obvious tendency for the CWS to display an overall warm bias when compared against IMMS. When the interpolation model was cross-validated in Section 4.5 the equivalent plot (Figure 4.24; note the change in y-axis scale) displayed no signs of systematic bias, suggesting it is the CWS observations that display an overall warm bias as supposed to an interpolation model that systematically under-predicts.

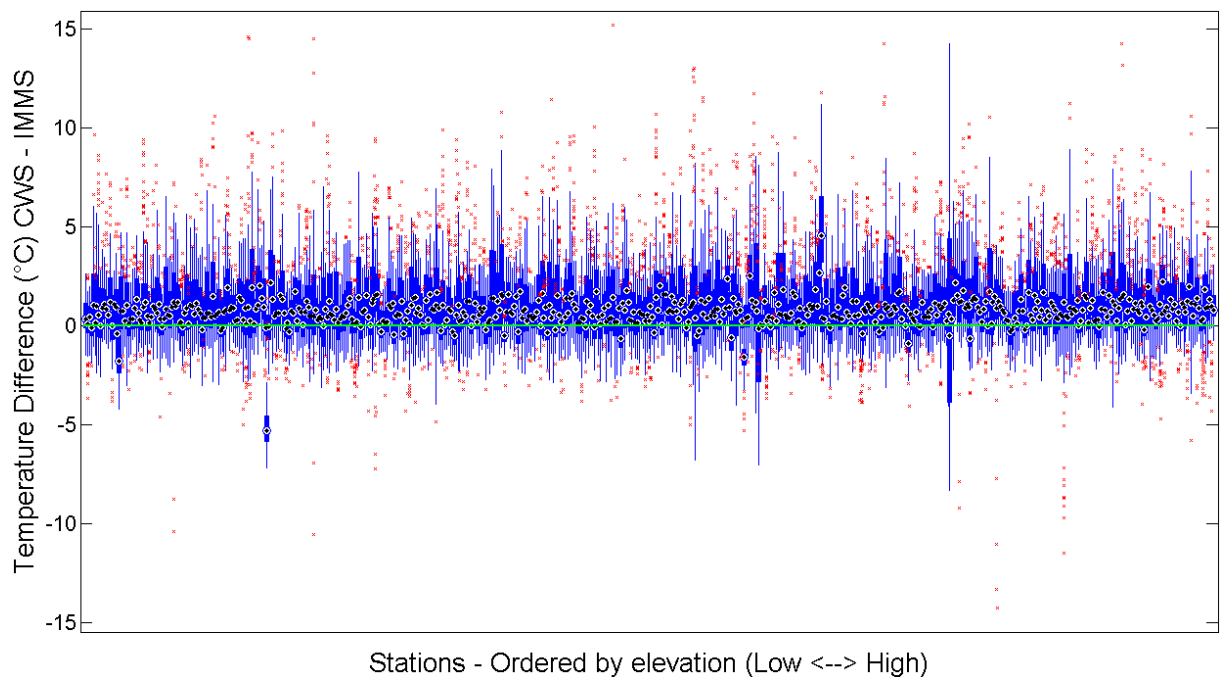


Figure 5.1. Boxplots of difference between the uncorrected CWS observations and IMMS; plotted individually for each CWS station over the summer period. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

In our field study the main cause of warm biases such as these were radiation-induced biases. It appears that this issue plagues real CWS data as well. Not all the stations shown in the box plot display such a warm tendency. This supports the notion that the warm biases seen in the other stations are radiation-induced as the field study highlighted that some station models, such as the VP2, display very little in the way of radiation-induced biases. This effect of station design is addressed further in Section 5.4. Note the presence of some very anomalous stations in the figure as well, for example the station whose interquartile range and whiskers all lie below 0 °C with a median temperature difference of -5.3 °C. On further investigation this particular CWS station displayed temperature observations that were consistently ~ 5 °C below the

IMMS estimate. This is probably due to a calibration bias, something our bias correction model attempts to correct for (Section 5.6).

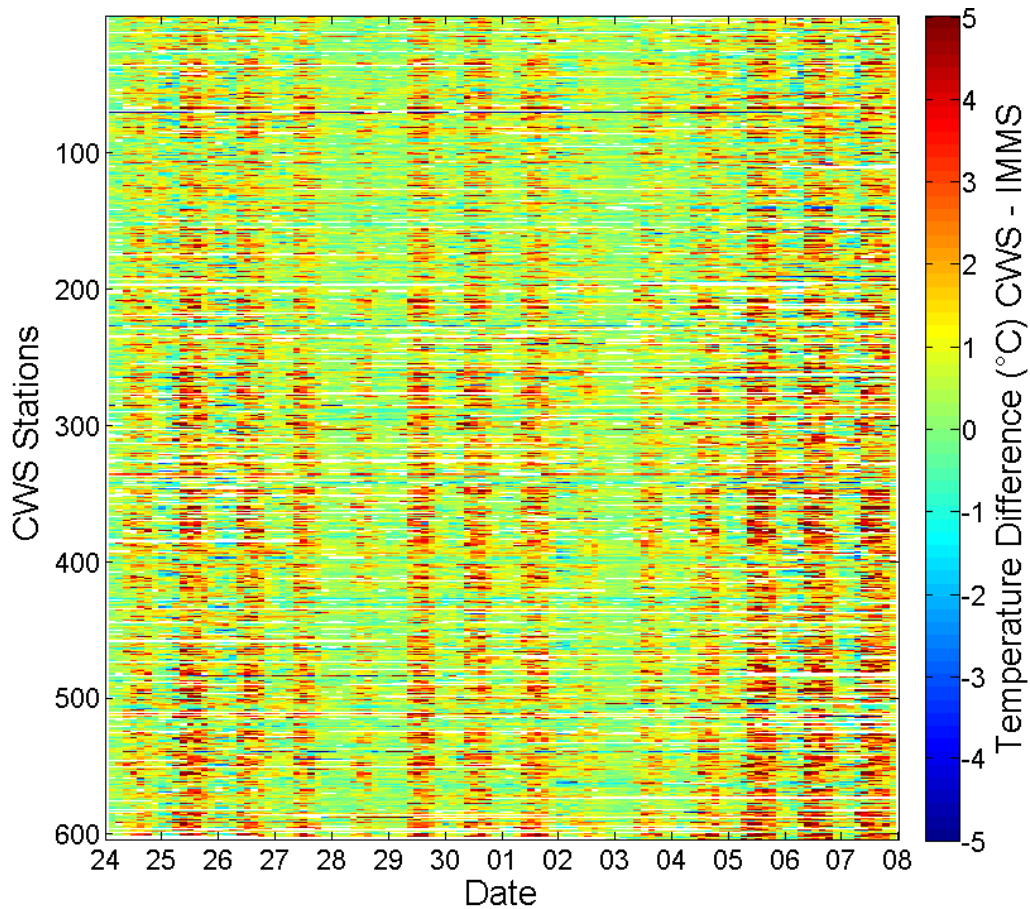


Figure 5.2. Visualisation of the difference between the uncorrected CWS observations and IMMS for each station (rows) and at each timestep (columns) over the summer period. Ticks on the x-axis indicate midnight at the start of that date.

There is a possibility that the warm bias may instead be a function of siting. For example there are a higher proportion of CWS stations located in urban and suburban areas than professional MMS stations (Section 5.5.2). As temperatures in urban areas tend to be higher than surrounding rural areas, thanks to the urban heat island effect, this could explain the warm tendency together with an ‘error’ in the interpolation model. However, Figure 5.2 poses a strong argument against this theory. Were the warm bias primarily a function of the urban heat island effect we would also expect to see an overall warm bias at night as well, but this is not evident. It is also important to note that the interpolation model should already account for changes in land cover, owing to the inclusion of the UKV and the predictor that represents urbanisation. Figure 5.2 shows a clear banding with a tendency for warm biases during the day and minimal temperature difference at night. On certain days the magnitude of the warm bias is not as significant; for example, compare the 28th against the 6th. In the satellite images for these days (Figure 5.3) there is an obvious difference in cloud cover, which

dramatically influences the strength of incoming radiation, and thus the degree of radiation-induced measurement bias.



Figure 5.3. Visible satellite images, from the Meteosat Second Generation satellite, during the summer period for the dates: a) 28th June 2013 12:00, b) 6th July 2013, 12:00. Source: BADC (badc.nerc.ac.uk).

As described in Section 2.1.4, WOW users can rate their site based on the quality of their thermometer readings, which in turn influences their sites overall star rating. Figure 5.4 and Figure 5.5 show that stations with different star and temperature ratings display slightly different 'CWS minus IMMS' statistics. Reassuringly the mean discrepancy is closer to zero for stations with the highest star and temperature ratings; with a tendency for a smaller standard deviation as well. This implies that such metadata ratings are likely to be informative of the degree of bias a station exhibits. A subject for future work would be to use such metadata ratings as a prior in the bias correction model to help quantify the expected bias and uncertainty.

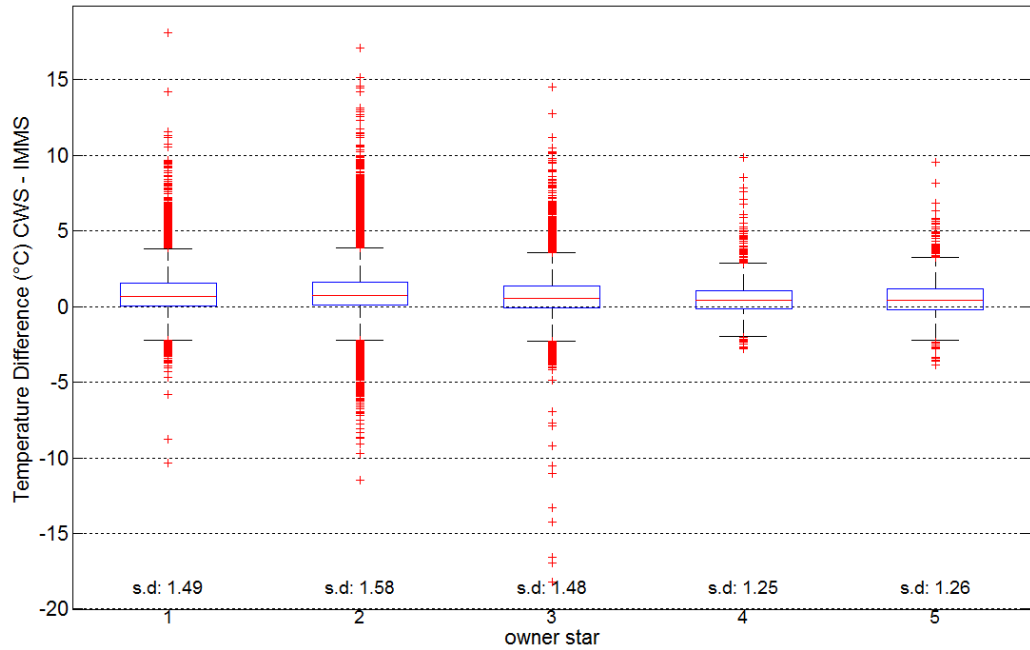


Figure 5.4. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed star rating (higher = better). Values at the bottom denote the standard deviation. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

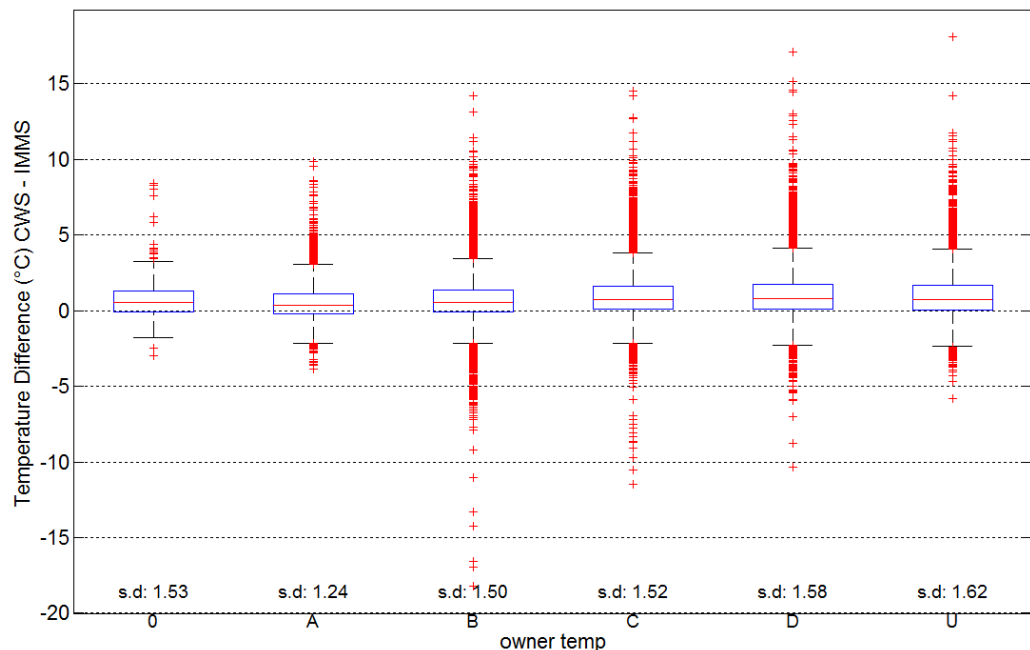


Figure 5.5. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed temperature rating. Ratings range from A, the highest quality, though to D the lowest. U denotes unknown quality and 0 implies a site with supposedly no temperature observations. Values at the bottom denote the standard deviation. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

This section has shown that there are significant differences between the CWS and IMMS. Although this section makes no attempt to quantify the relative causes of these differences, there is strong evidence that, as in the field study, real CWS data contains significant instrumental biases; most notably, radiation-induced biases. In the upcoming sections we detail approaches to learn these instrumental biases so that they may be corrected for.

5.3. Addressing radiation bias

Both the intercomparison field study (Section 3.2) and the exploratory analysis of WOW data (Section 5.2) clearly indicate that many CWS exhibit large temperature biases with a strong dependency on the strength of incoming solar radiation. Section 3.2 also demonstrated that with an accurate estimate of the Global Horizontal Irradiance (GHI) at a CWS location we can successfully correct these radiation-induced temperature biases and lower the residual variance. In practise however – out of the thousands of CWS very few are collocated with an accurate pyranometer. In fact there are only 80 or so MMS sites measuring GHI across Great Britain (Figure 5.6).

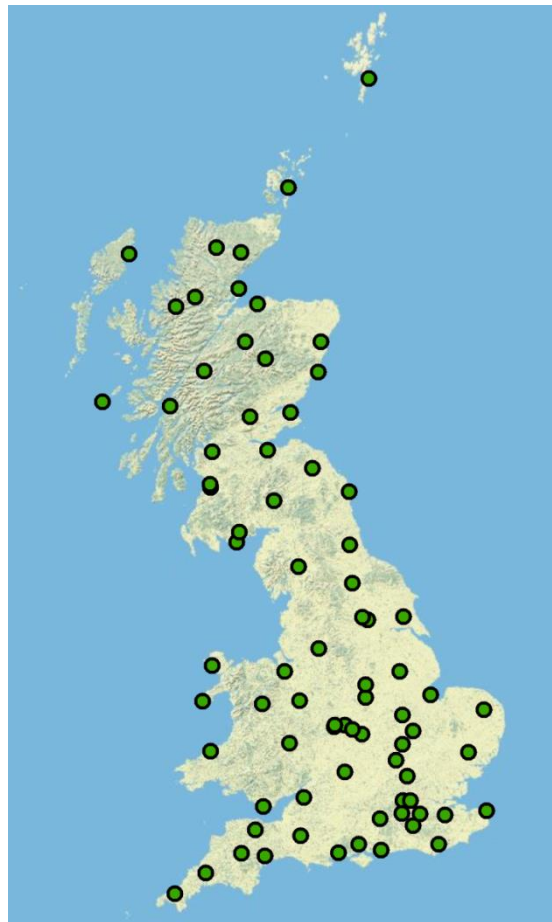


Figure 5.6. Spatial distribution of Met Office MMS stations that regularly record Global Horizontal Irradiation (GHI).

To overcome this problem, the Bayesian linear regression model, as used in Section 4.3, is implemented here to interpolate GHI measurements made at MMS sites, to every CWS location. These estimates can then be fed into the bias correction model as shown in Figure 1.2. At a given timestep the strength of the GHI varies spatially across the country, influenced by several factors. Covariates of GHI are used as predictors in the interpolation model. We must be able to estimate reliably these covariates at any location in Great Britain. The first covariate is an estimate of the theoretical GHI under perfect clear-sky conditions. Obviously the time of day/year influences GHI through changes in the solar zenith angle. From this angle it is possible to derive the clear-sky GHI, as explained further in Section 5.3.1. As the sky is rarely completely clear across Great Britain, visible and infrared satellite images of cloud cover are also used to account for the discrepancy between the clear-sky GHI and the amount of solar energy that actually reaches the surface. The imagery used and how it was processed is detailed in Section 5.3.2. Section 5.3.3 details exactly how these covariates are fed into the Bayesian regression model and Section 5.3.4 assesses the performance of the model.

The approach implemented here is an alternative to pre-existing services that provide, at a cost, surface solar irradiance estimates. Most notable is the HelioClim Project (Blanc, et al., 2011) and its HelioClim-3 dataset, which provides irradiance values at 3 km spatial resolution and 15 min temporal resolution over Europe, Africa and the Atlantic. It too exploits Meteosat Second Generation satellite images, but uses the Heliosat-2 method (Rigollier, et al., 2004) to interpret the images. This is run in near-real time. Their approach focuses on calculating a *cloud index* to quantify the difference between what the satellite observed and what should be observed over that pixel were the sky clear. By incorporating and interpolating surface observations we take a somewhat different approach, as detailed in Section 5.3.3. Although we employ a Bayesian regression model to perform the interpolation, many other techniques have also been demonstrated, albeit mainly for daily observations. These include thin plate splines (Xia, et al., (2000); Jeffrey, et al., (2001)), artificial neural networks (Bosch, et al., 2008) and regression kriging (Gutierrez-Corea, et al., 2014). Our aim was not to compare our approach against these alternative methods, but merely to produce reliable GHI estimates.

5.3.1. Clear-sky global horizontal irradiance

Clear-sky GHI is a measure of the sun's power (W m^{-2}) at a given surface location. It assumes zero cloud cover and therefore represents the maximum possible power that can be expected. This makes it a useful predictor of actual GHI when combined with

cloud cover data. There are several approaches for calculating the clear-sky GHI (Reno, et al., 2012). Here the simplistic Robledo-Soler algorithm is used (Robledo & Soler, 2000):

$$GHI = 1159.24(\cos z)^{1.179} \exp(-0.0019(90^\circ - z)) \quad (12)$$

This model is purely geometric, i.e. the clear-sky GHI is merely a function of the solar zenith angle, z , and does not incorporate any meteorological data. Improved estimates may come from using a model that does integrate such data; e.g. the McClear model (Lefèvre, et al., 2013), which exploits aerosol, water vapour and ozone information exported by the MACC project. As illustrated in Figure 5.7, the difference between the two estimates tends to be relatively small. Given that our interpolation model is anchored by pyranometer observations we assume the impact of this difference is negligible. Due to Equation (12)'s simplicity, and to avoid purchasing McClear data, the Robledo-Soler approach was favoured here instead.

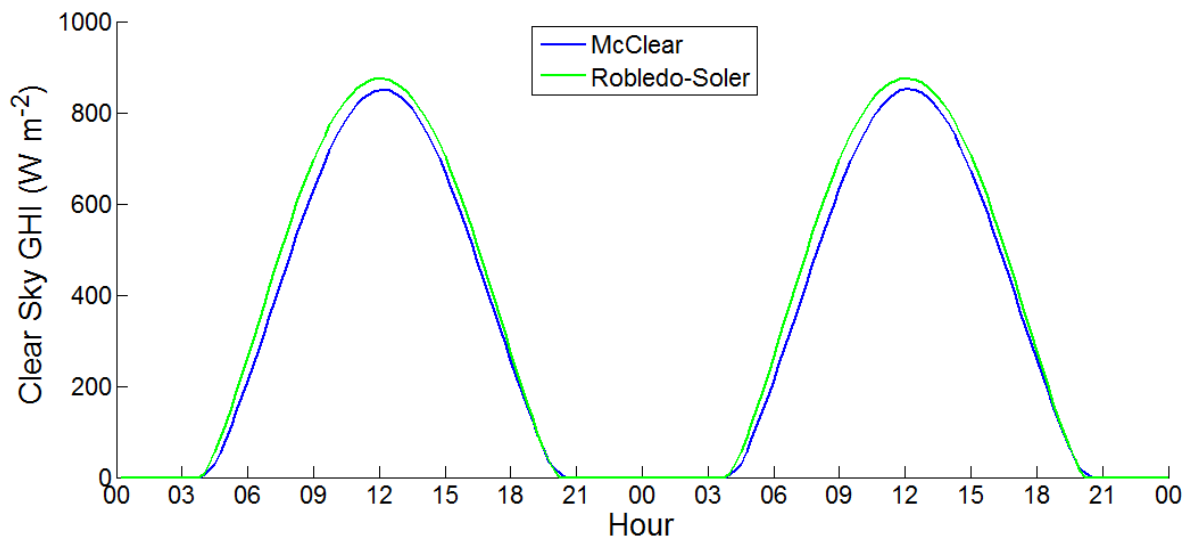


Figure 5.7. Comparison of the clear-sky GHI estimates from two different approaches over the 1st to 2nd of June 2004.

The zenith angle, fed into the Robledo-Soler equation, was calculated using a MATLAB implementation of the algorithm presented by Reda & Andreas (2004). Inputs to this equation are the location's latitude, longitude and altitude as well as the time and date.

5.3.2. Satellite imagery

Cloud cover changes are responsible for the high spatial and temporal variability within the GHI field. The variability in cloud cover is quantified using infrared and visible satellite images. The images come from the Meteosat Second Generation (MSG)

geostationary satellites operated by EUMETSAT, which have a resolution of ~ 1 km. The images were reprojected using a series of control points with ArcGIS' georeferencing tool to a 700×1300 grid such that it aligned with a 1 km resolution British national grid projection (Figure 5.8). The green country outlines added by EUMETSAT were replaced by the nearest 'non-green' pixel. The archived satellite images are freely available from the BADC website (badc.nerc.ac.uk).

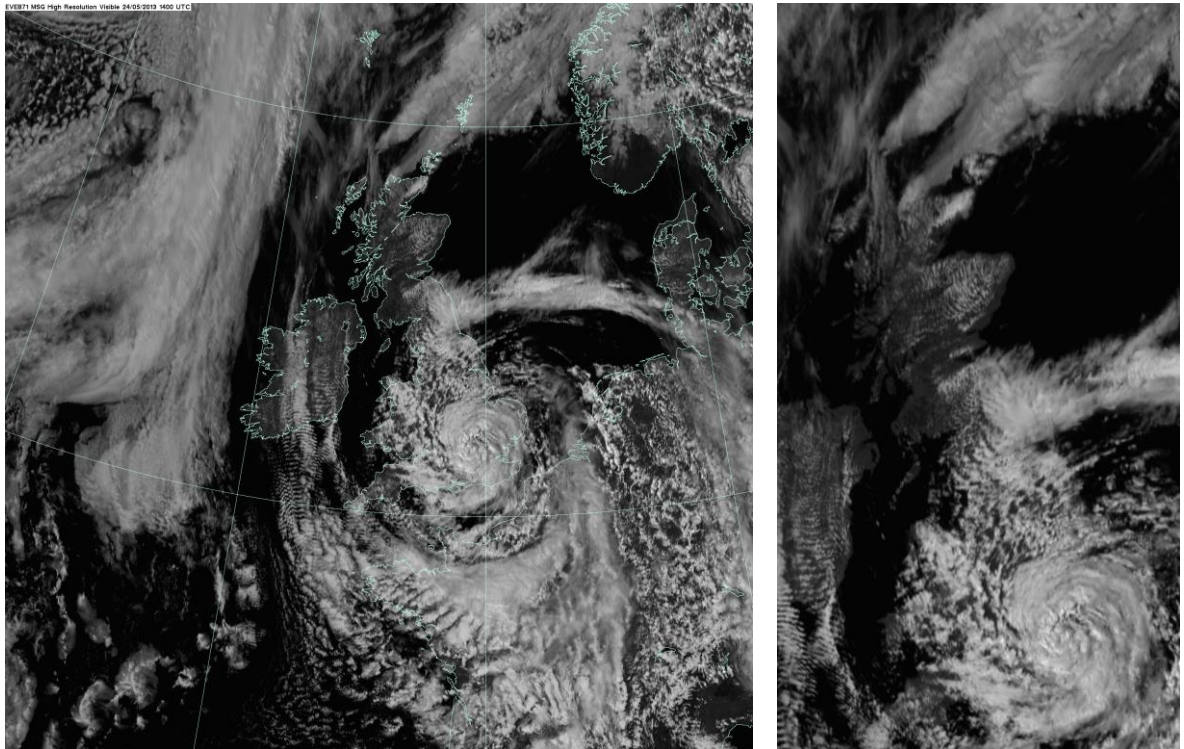


Figure 5.8. Visible MSG satellite image of Great Britain on 24th May 2013 at 14:00 GMT before (left) and after (right) removing the green country outlines and reprojecting to the British national grid.

Once in the correct projection, the pixel values could be extracted at the locations of the stations. For a given location, the 25 nearest pixel values were selected to represent this location. By simply selecting the nearest pixels the pattern of those selected isn't necessarily a 5×5 grid. Pixels over land and water are treated equally. This number of pixels is used, in part, to account for errors in the spatial reprojection, but also it means that clouds that may have recently passed over the station are sampled. This number also produces low cross-validation errors. The average of these pixels was taken, having assigned equal weight to each pixel. As an alternative, the pixels were also weighted by distance to the station location, but as this gave no improvement to the cross-validation error the first approach was favoured for simplicity. Overall this approach is relatively simplistic, future work may wish to more accurately consider the field of view of a pyranometer, and incorporate each

pyranometer's sky view factor. The direction and speed of cloud movement should also be accounted for.

Many other modifications of the satellite data were tested, for example finding the lowest value of each pixel over the period and using this to infer what the pixel value would be without cloud. For the infrared images this was also performed for each hour individually to account for the diurnal temperature fluctuations. The difference from these minimum values was then used to infer how cloudy the pixel was. However, the improvement in cross-validation error gained by using these as predictors was minimal, if present at all, and due to the extra computational cost required they were omitted from the final model. The variance of the 25 pixels was also calculated to act as a proxy for how scattered the clouds were, but likewise this yielded no improvement and was omitted.

If these satellite images are to be used to correct CWS radiation biases operationally in near real-time, i.e. so the corrected observations can be fed into a data assimilation cycle, then the images would need to be obtained and processed in near real-time. It is therefore beneficial to keep the processing of these images as simple as possible. Forecasting agencies, such as the Met Office, already assimilate satellite imagery into their NWP (Joo, et al., 2013); therefore data accessibility should not be an issue.

It is important to note that these images were only publically available at hourly resolution; valid *on* the hour. Because we rely on these satellite images to perform the interpolation our interpolated estimates are also valid *on* the hour. This leads to a 10 minute discrepancy between our MMS & CWS temperature observations and our radiation estimates. The impact of this on the overall temperature bias estimates is assumed to be acceptably small, with the temporal averaging introduced below and the spatial averaging of the satellite imagery helping to lessen its impact.

5.3.3. Model structure

The interpolation model used here to interpolate GHI has exactly the same form as the model used to interpolate temperature (Section 4.3). It too allows the regression coefficients and model uncertainties to propagate through time, updated iteratively at each timestep. Propagating the regression coefficients like this ensures they exhibit greater temporal continuity, leading to a more stable model. This leads to a slight drop in model error over learning the regression coefficients from fresh at each timestep.

As explained in Section 3.2 the relationship between the global radiation and the temperature bias was often stronger when, instead of simply using the point radiation

observations at the equivalent timestep, past radiation observations were weighted with an exponential function:

$$\int_0^{60} e^{-\lambda x} dx = 1 \quad (13)$$

As equation (13) shows a weighting is calculated for each minute radiation observation, with x the number of minutes from the time of the temperature observation ranging from 0 to 60 minutes beforehand. Here λ is set as 0.04 to achieve the curve show in Figure 3.10. Note that the weightings are normalised to sum to 1.

As the aim of this radiation interpolation model is to better predict the temperature bias for a given CWS, it is logical to interpolate these temporally smoothed radiation values rather than the point observations. Not only does it have a stronger relationship, but using these smoothed radiation values improves the accuracy of interpolation model (Figure 5.9). This is likely a result of smoothing out the often noisy point global radiation observations.

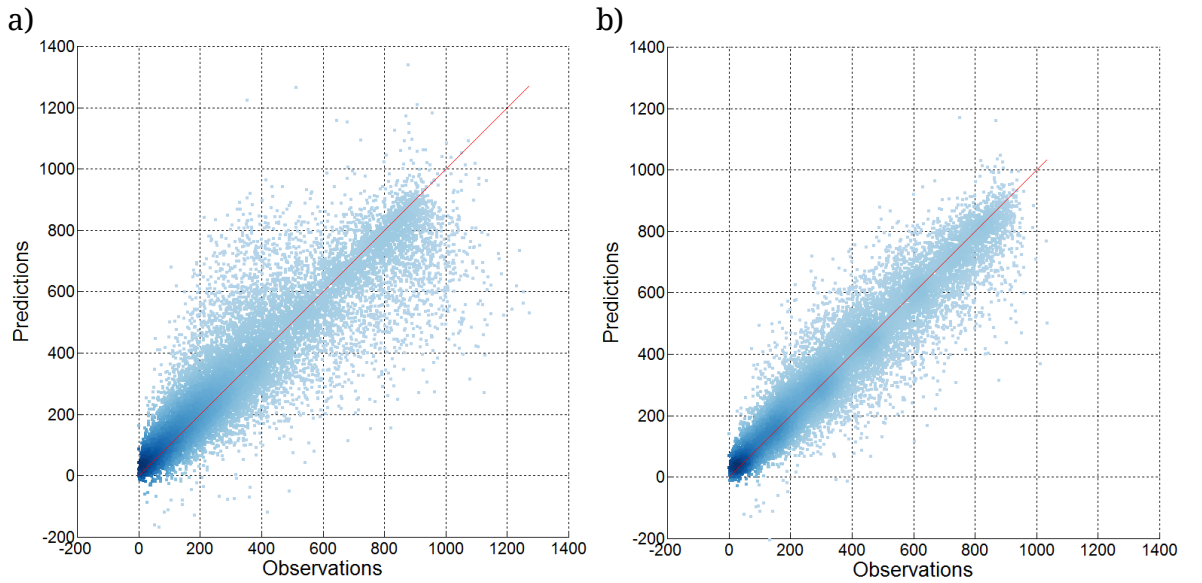


Figure 5.9. 1:1 plots of radiation interpolation model predictions verified against MMS global radiation observations using 10-fold cross-validation. In a) point observations are used, whereas in b) observations over the past hour have been exponentially weighted. Only data from the summer period is shown. The deeper the colour the higher the density of points.

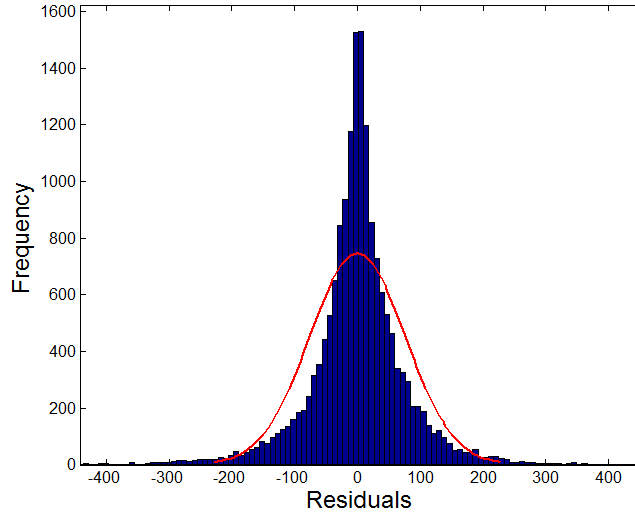


Figure 5.10. Histogram of residuals from predictions made by the radiation interpolation model when using exponentially weighted global radiation observations as the target, verified against MMS observations using 10-fold cross-validation. Only data from the summer period is shown. The red line represents a Gaussian distribution fitted to the residuals.

Even when the radiation values were temporally smoothed in this way the model still had issues. The data points were not well distributed across the domain (Figure 5.9) nor were the residuals close to displaying a Gaussian distribution (Figure 5.10); instead it strongly resembles a t-distribution. The solution was to take the log of the smoothed radiation values first. However, taking the log of values close to 0 can cause issues so a constant value of 100 W m^{-2} was added to every smoothed radiation value before taking the log. Therefore, the final radiation variable to be interpolated was as follows.

$$l_{Rad} = \log(w_{Rad} + 100) \quad (14)$$

Where w_{Rad} is the point minute resolution radiation observations over the proceeding hour, weighted exponentially using the approach shown in Equation (13). As evident in next section, where the model's performance is assessed, using the transformation to l_{Rad} dramatically improves the spread of points across the domain (Figure 5.11) and ensures the residuals more closely resemble a Gaussian distribution (Figure 5.12). There was a concern that this log transformation may weaken the relationship with CWS temperature bias. To test this, the same transformation was applied to our field study observations. As Figure 3.11 and Figure 3.12 show there was minimal effect to the strength of the relationship when this log transformation was applied. The final result is a target variable representative of the strength of incoming solar radiation, which is a reliable predictor of CWS temperature bias, and can be accurately interpolated to CWS locations.

As previously mentioned the model relies on three key predictors. A clear-sky GHI estimate and estimates of cloud cover derived from the visible and infrared satellite images. Each of these sets of values form both 1st and 2nd order basis functions within the design matrix, X (Equation (1)). The design matrix also includes interaction terms for clear-sky GHI with the visible and infrared image values separately. As the target variable, l_{Rad} , underwent an exponential weighting and log transformation it was also logical to perform the same transformations to the clear-sky GHI estimates as well. Unlike temperature interpolation, adding RBFs led to little improvement; in fact at times they dramatically increased the error. For this reason they were omitted.

Depending on the strength of l_{Rad} the bias correction model (Section 5.6) performs different actions. It is therefore imperative that the interpolation model can produce an estimate of l_{Rad} at every timestep. However, when a visible image contains partial darkness it cannot be used. To overcome this challenge the interpolation model must also be run using just the clear-sky GHI and infrared image. As becomes evident in the next section, running the model without the visible image is not as accurate. However, it is still acceptable, and in future the attached uncertainty estimate could be used to account for the reduced confidence.

5.3.4. Model performance

The radiation interpolation model performed satisfactorily when verified against the MMS observations using 10-fold cross-validation. Figure 5.11 shows that, thanks to the log transformation, the points fall relatively uniformly across the domain, and Figure 5.12 shows that the residuals are closer approximated by a Gaussian. The residuals are also unbiased with a zero mean. Admittedly the distribution of the residuals is still closer to that of a t-distribution, which allows for heavier tails, however to keep our model computationally convenient and efficient we stayed in the class of Gaussian distributions.

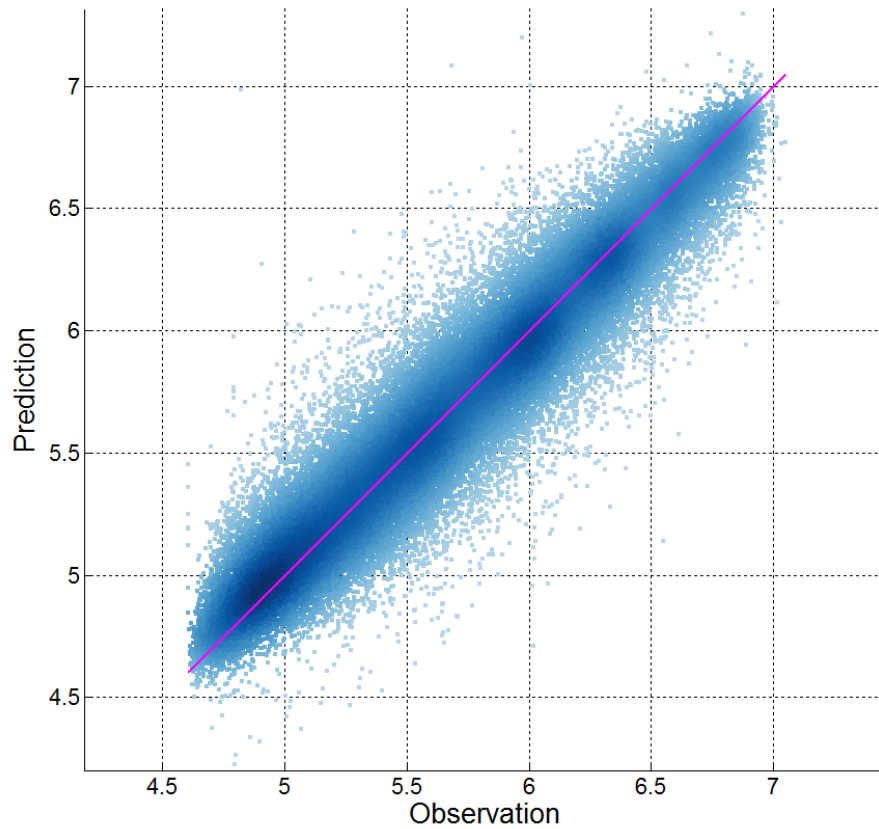


Figure 5.11. 1:1 plot of predictions from the final radiation interpolation model against MMS observations when verified using 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.

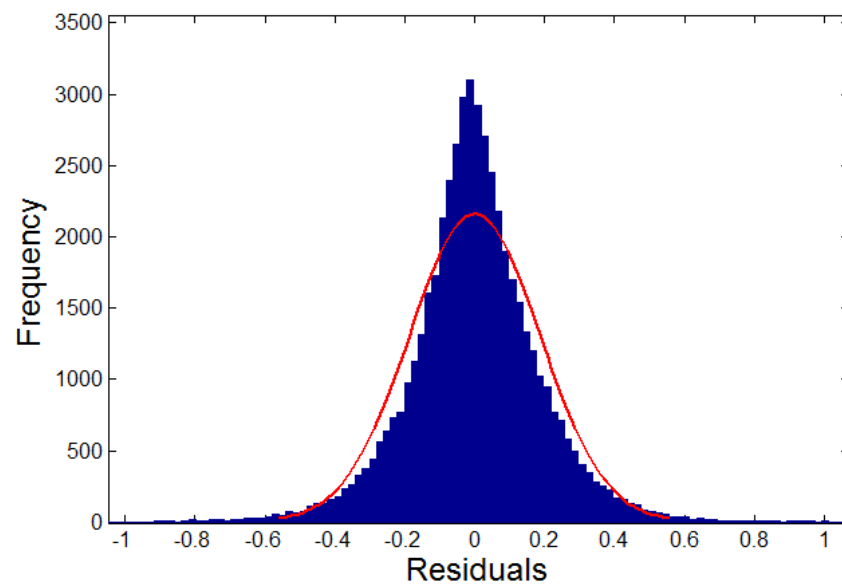


Figure 5.12. Residuals from the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The Model includes both visible and infrared satellite imagery.

Each prediction is a distribution with a mean and variance. It is important that the model is validated probabilistically to ensure that not only is the mean estimate

accurate but that the variance fairly represents the uncertainty in the prediction. Getting these uncertainties correct is important as in future work we wish to propagate them through to the bias correction model to influence the uncertainty in our bias corrections. In the z-score plot (Figure 5.13) we would expect to see 95% of points fall between ± 2 indicating a good probabilistic model. Fortunately our model approximately shows this.

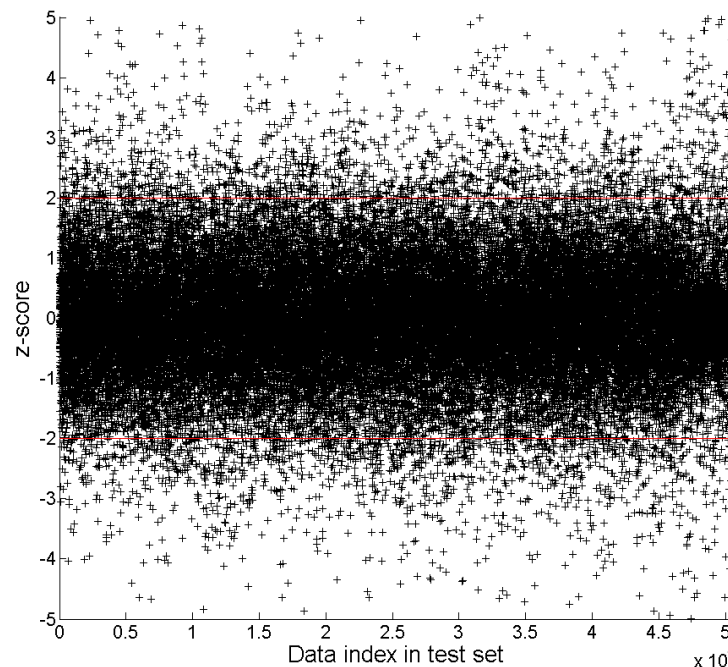


Figure 5.13 Z-scores plot for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.

Also as the points in the coverage plot (Figure 5.14) fall close to the red 1:1 line and the rank histogram (Figure 5.15) is acceptably flat we can be confident that our model is not practically significantly over- or underconfident about the predictions it makes.

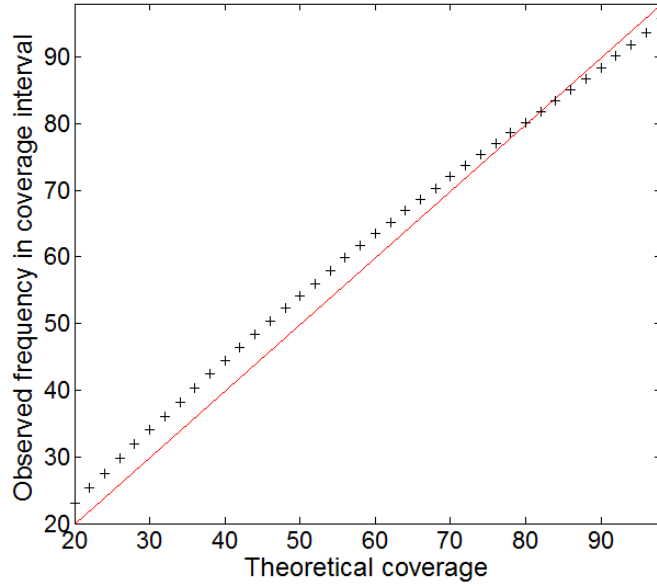


Figure 5.14. Coverage plot for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery. It plots the theoretical centred confidence interval against the observed frequency.

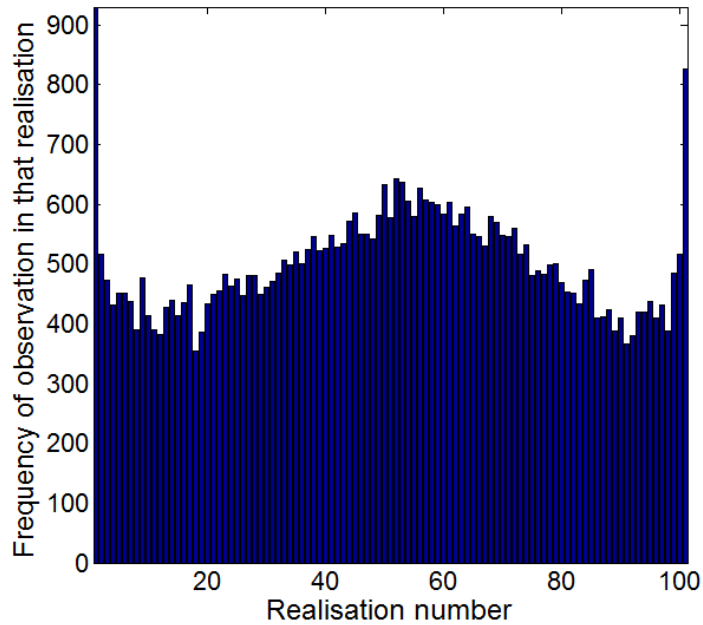


Figure 5.15. Rank Histogram (Hamill, 2001) for the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes both visible and infrared satellite imagery.

When the visible imagery is removed from the model, the error increases, as indicated by the greater spread about the 1:1 line in Figure 5.16. However there is still a good spread across the domain and the residuals are relatively well approximated by a Gaussian (Figure 5.17). It therefore provides an adequate solution for estimating l_{Rad} when visible images are unavailable.

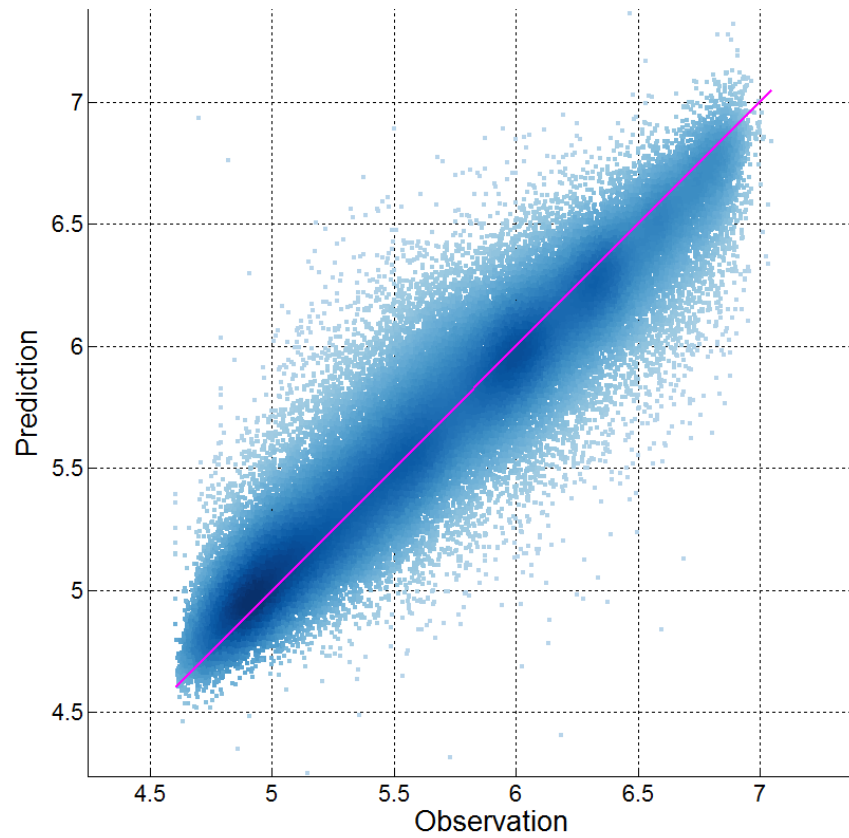


Figure 5.16. 1:1 plot of predictions from the radiation interpolation model against MMS observations when verified with 10-fold cross-validation. All four periods are included in this plot. The model includes infrared satellite imagery, but not visible.

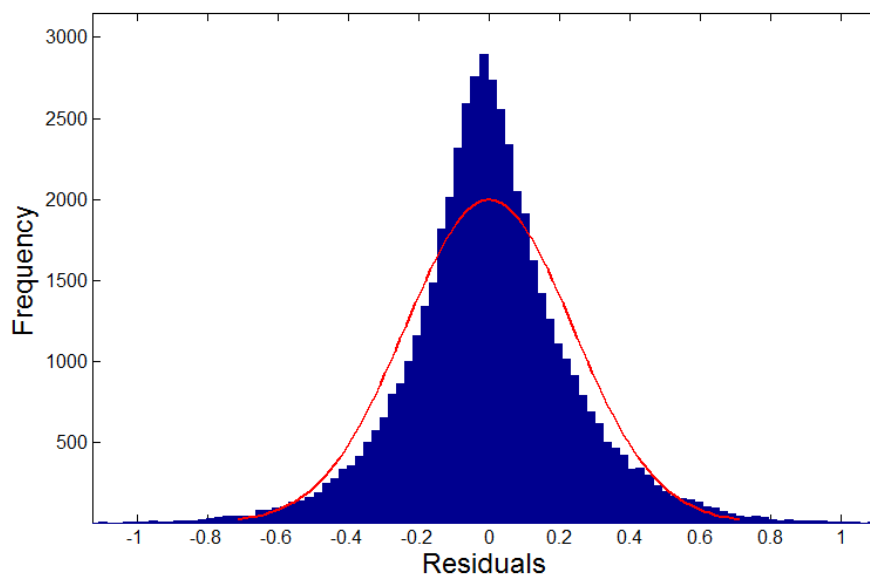


Figure 5.17. Residuals from the radiation interpolation model when verified against MMS observations with 10-fold cross-validation. All four periods are included in this plot. The model includes infrared satellite imagery, but not visible.

5.4. Addressing station design

Every CWS model has a different design. Figure 1.1 illustrates the variety of different designs just within field study stations. This field study (Section 3.2) was vital for highlighting the significant effect the design plays in determining the magnitude of the instrumental biases. As such, when dealing with citizen data submitted to websites such as WOW, knowing the model of the station can provide *a priori* information about the type and magnitude of bias we would expect that station to exhibit. Below we detail an approach for automatically extracting the model name from the user-contributed metadata on WOW. Then we propose that out of the dozens of models available, there exist several subsets of stations, each with a common style of radiation shielding (Section 5.4.2). It is this shielding that encases the CWS thermistor and appears to play a key role in determining the degree of radiation-induced biases exhibited. Therefore if our model can establish what type of design a station is, it stands a better chance of estimating the radiation-induced bias. Section 5.4.3 shows signs of this effect of design type in real CWS data from WOW.

5.4.1. Automatic extraction of model name from metadata

As the bias correction model is tested with real CWS data from WOW, we would like to obtain the model name of each WOW station. However, as detailed in Section 2.3.4, listing the manufacturer and model of your station is not compulsory when signing up to WOW; nor is there a consistent list of stations to choose from. The only way of finding the station type is from the textual metadata a user may choose to write about their station in the sections titled *Site Description* and *Additional Information*. Doing this manually is very time-consuming. However scanning the text computationally to identify keywords proved a viable alternative. The following steps were taken to semi-automatically extract the model names from these textual metadata boxes:

1. An initial dictionary of keywords was created. For example the keyword ‘VP2’ would be indicative of the ‘Davis Vantage Pro2’ station model.
2. Certain keywords were also associated with ‘enabler’ keywords, these enabler words must also exist for the station to be given a particular model name. For example the word ‘Pro’ alone does not automatically mean the station is a ‘Davis Vantage Pro’, the enabler words ‘Davis’ or ‘Vantage’ must also be present.
3. Keywords may also be associated with ‘denier’ keywords, which prevent a station being assigned to a model name. For example, a keyword of ‘VP2’ would be denied from being labelled as naturally aspirated ‘Davis Vantage Pro

- 2' if the denier word 'FARS' is present. As this would imply the station is actually aspirated and thus should be assigned as a 'Davis Vantage Pro 2 FARS'.
4. The textual metadata was web scraped from the WOW website using a script written in the programming language *Ruby* and passed through the dictionary of keywords looking for any matches. Once a match was made and the station passed the enabler and denier caveats then the station was designated as a particular model.
5. Stations with no metadata were designated as *Unknown*. Stations with metadata, but no matches were then manually inspected to check for possible keywords missing from the dictionary. If any were spotted then the dictionary was updated and the process rerun.

This approach successfully assigned a model name to just over 40% of the WOW stations used. The low percentage is firstly because only 70% of users actually filled in at least one of text boxes, and secondly this keyword approach could only find a model name in just 60% of these. The stations for which a model name could be assigned is now categorised into one of the more general shielding design classes as explained in the next section.

5.4.2. Design classes

Radiation shielding design plays a dominant role in determining the degree of radiation-induced biases exhibited by a station. The field study results (Section 3.2) show that models with a poor shield design can display warm biases well over 5 °C under strong insolation. Here we propose 7 key shielding designs, each of which displays a different degree of bias, into which all station models uploading to WOW can be categorised.

Table 4 lists the 7 design classes. They range from the professional standard – the Stevenson screen – through various louvered designs with varying degrees of shielding and ventilation performance – into more encased designs which are prone to overheating because air struggles to flow freely around the thermistor.

Table 4. A list of the 7 design classes to which a CWS can be allocated based upon the style of radiation shielding and the subsequent radiation-induced biases exhibited.

<p>Stevenson Screen</p>  <p>e.g. Wooden or plastic Stevenson screens. Including international equivalents of the UK Stevenson screen.</p>	<p>Aspirated</p>  <p>e.g. Davis VP2 FARS.</p>	<p>Quality Louvered</p>  <p>e.g. Davis VP2, Davis Vantage Pro, Vaisala WXT520.</p>	<p>Underslung Louvered</p>  <p>e.g. Davis Vantage Vue, Am. Weather WS-2090, Fine Offset HP1000.</p>
<p>Encased Louvered</p>  <p>e.g. Fine Offset WH1080, OS WMR100, La Crosse WS28** series.</p>	<p>Encased</p>  <p>e.g. OS WMR200, OS WMR918.</p>	<p>Encased Extreme</p>  <p>e.g. La Crosse WS2350, La Crosse 25** series, TFA 35.1095, OS LW301.</p>	

The choice of these 7 categories and the biases we would expect to see from each have been informed by the field study results. Figure 5.18 illustrates the temperature bias as a function of global radiation for each of these 7 design classes as learnt in the field trial. Although here the relationship is shown with global radiation, in practice l_{Rad} is used, as this is what we interpolate to CWS locations. The update step from the Bayesian linear regression model (Section 4.3.2) was used to learn the regression coefficients that represent this relationship. The design matrix incorporated both first and second order basis functions for l_{Rad} , and a constant term. The addition of the second order term helps improve the model fit (Figure 5.19). This Bayesian regression approach also estimates a covariance matrix for the regression coefficients; later in the model this is used to ensure we propagate our uncertainty.

Although the field study showed signs (e.g. Figure 3.7) that the wind speed also influences the temperature bias, its influence was minimal compared to the effect of radiation. For simplicity it is therefore not used as a predictor.

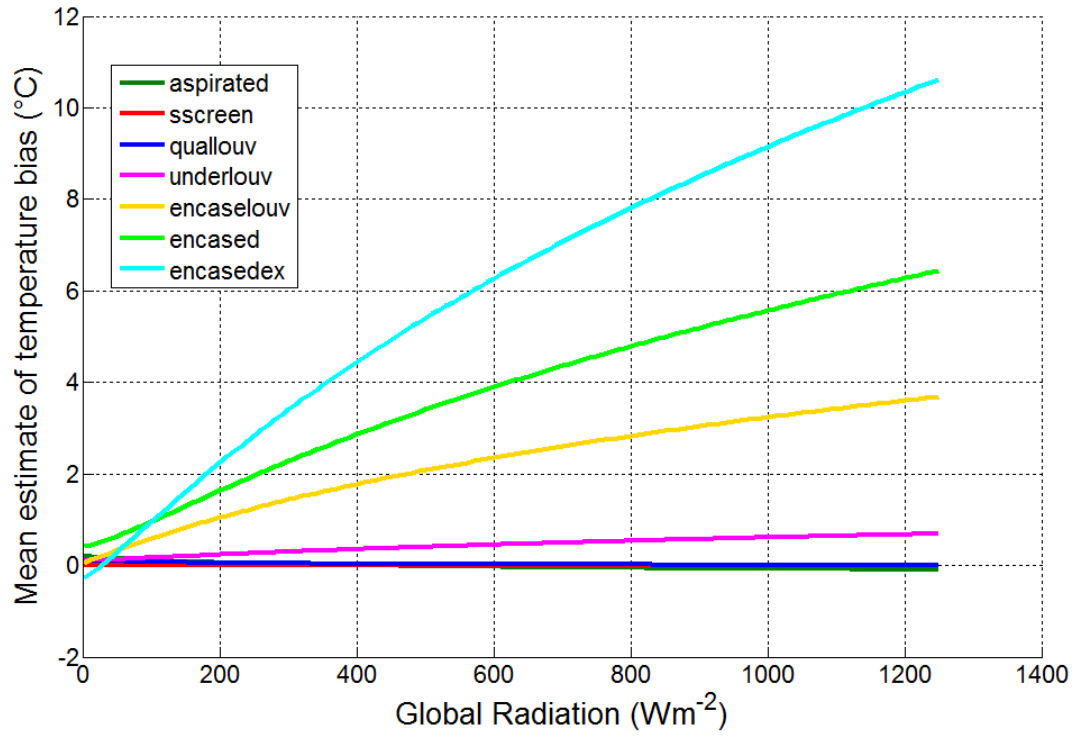


Figure 5.18. Relationship between global radiation and the temperature bias for each of the 7 design classes as learnt from the intercomparison field study. The temperature bias is estimated based upon the equivalent l_{Rad} value for global radiation values of 0 W m^{-2} through to 1250 W m^{-2} at 1 W m^{-2} intervals. A second order regression function is used to predict the temperature bias from the equivalent l_{Rad} values using the regression coefficient mean terms μ_β learnt from field study for CWS belonging to the given class.

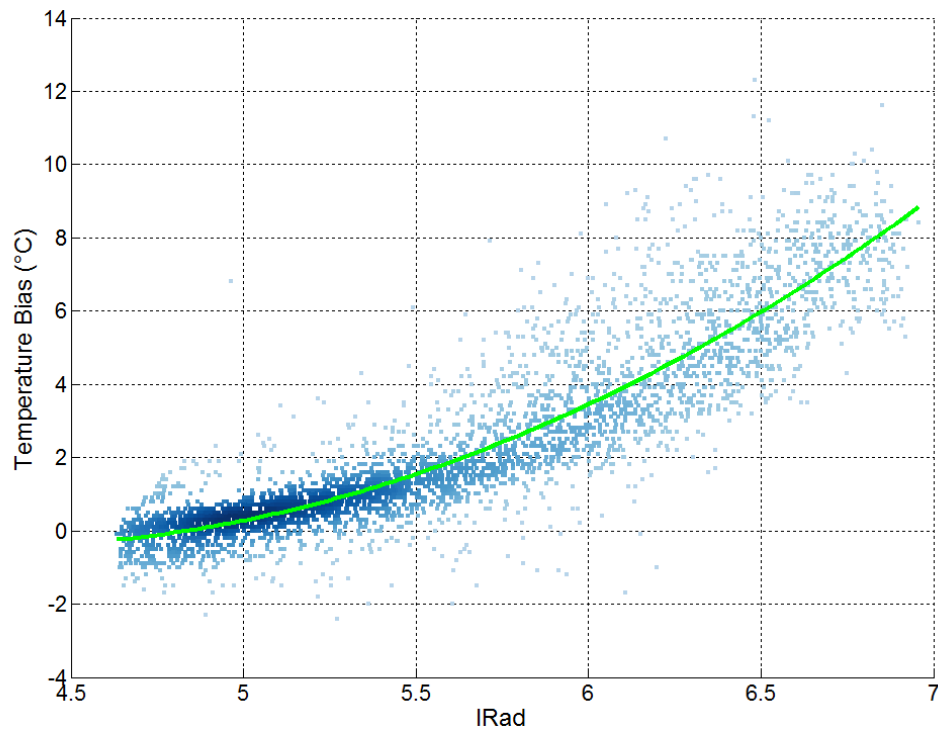


Figure 5.19. Relationship between La Crosse WS2350 temperature bias and radiation (log transformed to l_{Rad}). Green line indicates the fitted model used to represent stations belonging to the 'Extreme Encased' design class.

As the models chosen for the field study were selected because they are the most popular this ensures that a large proportion of the models used on WOW have been tested and can therefore be accurately allocated to one of these design classes. Other models used on WOW were manually allocated to a particular class based simply on their design characteristics. For stations we are unsure about, the model allows them to belong to more than one class, and for stations without metadata we allocate an equal probability to every class as explained in more detail in Section 5.6. If we accidentally assign a CWS to the wrong design class, i.e. it does not display that particular class' bias characteristics, then thanks to the Bayesian updating approach detailed below (Section 5.6) the station is not confined to that class and can gradually shift to a more appropriate class. Given that at any point a citizen may decide to swap their station for one with a different design, perhaps without updating their metadata, this Bayesian approach provides an elegant solution for handling such a change; i.e. if the new station exhibits different bias characteristics then its design class membership simply changes so that a more suitable bias correction is applied.

In Section 3.2 we noted that nearby obstructions can shadow a CWS from direct sunlight, thus affecting the magnitude of the radiation-induced biases it exhibits. Such an effect could cause a station to be allocated to the wrong class, particularly in winter when the sun is lower in the sky and thus shadowing would be more common. In order to better assess this issue a longer time period than the 2 week case study periods used later in Section 5.7 is required. If shadowing effects are a factor for a given CWS then we may see seasonal changes in its assigned design classes.

5.4.3. Evidence of design effect in WOW data

The field study showed evidence that different weather station design classes induce different degrees of bias. It is crucial however to check that real CWS stations allocated to a given design class exhibit biases characteristic to the design class to which they are assigned. As in Section 5.2 CWS data from WOW is used and compared against IMMS. The discrepancy between the two gives a sense of the bias. However it is important to reiterate that these discrepancies can result from not only radiation-induced biases, but also calibration biases, model error, representativity error and natural spatial variations.

Figure 5.20 shows box plots of the discrepancy with each box including stations whose metadata assigned them to one of the listed design classes. Fortunately the observed discrepancies agree relatively well with the biases we would expect from that particular class. For example the *Aspirated*, *Stevenson screen*, and *Quality Louvered* classes show the smallest discrepancies whereas the *Encased*, *Encased Louvered* and

Encased Extreme show the largest. The spread of the design classes is also larger for these latter design classes, which suggests we should place less confidence in the accuracy of their observations. We would expect to see the discrepancy of the *Encased Louvered* to be lower than the *Encased* and *Encased Extreme* classes, the fact it isn't implies one of the previously listed alternative explanations for the discrepancy is masking the relationship, or there may be issues due to small sample sizes within some of the classes, or because some stations have been wrongly assigned to a particular class. The figure presents data for the summer period only, selected as the difference between the classes appears most pronounced, however the other 3 case study periods show a similar pattern.

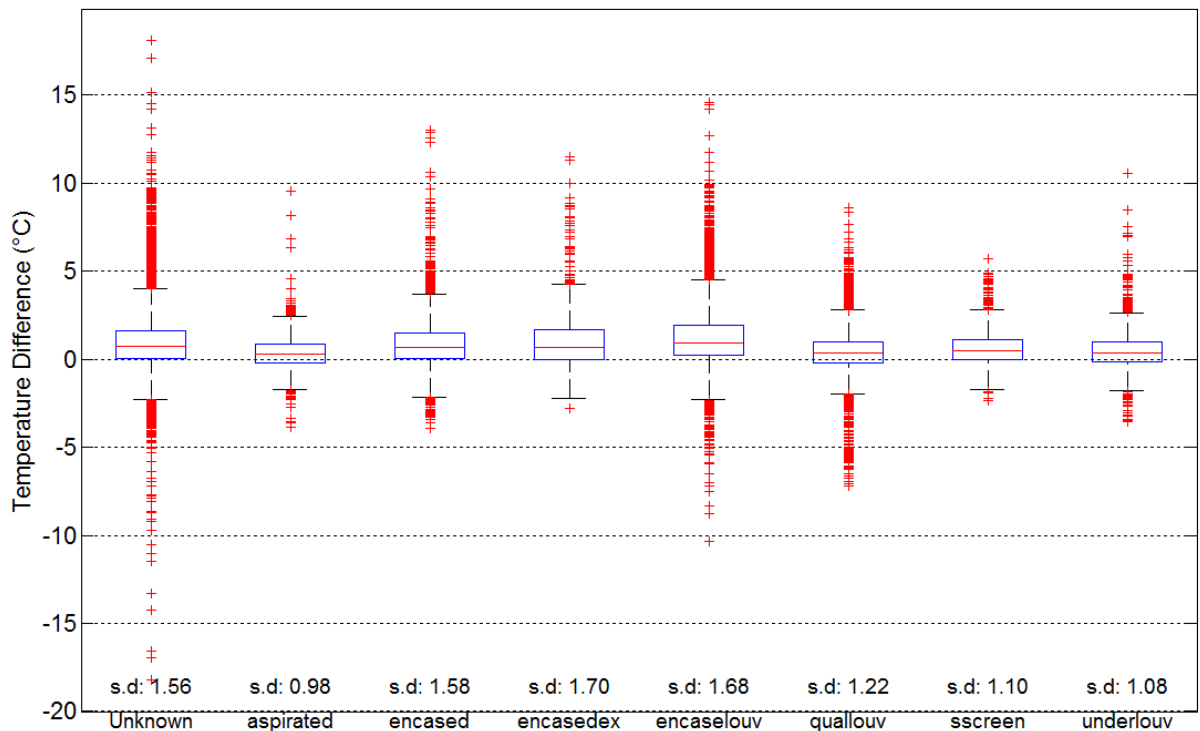


Figure 5.20. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their user-contributed model name and therefore designated design class. All available WOW stations (604 stations) are including using 3 hourly data throughout the summer period. Values at the bottom denote the standard deviation. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

Figure 5.21 gives a clearer sense of whether the biases we see in the real CWS data for a given design class correspond well with what we would expect; i.e. whether they closely resemble the biases seen during the intercomparison field study on which the design class characteristics are based (Figure 5.18). For the *Encased Louvered* class (Figure 5.21b) the observations agree well with what we would expect. For the *Quality Louvered* class (Figure 5.21a) the CWS observations appear systematically warmer

than we would expect when l_{Rad} is at its highest. This could suggest that the *Quality Louvered* station used in the field study is not representative of stations with a similar design type on WOW, or that other factors are causing the warm bias, or perhaps that for some stations the station type detailed within the metadata does not represent the bias characteristics it displays and therefore justifies allowing the station to change design class over time.

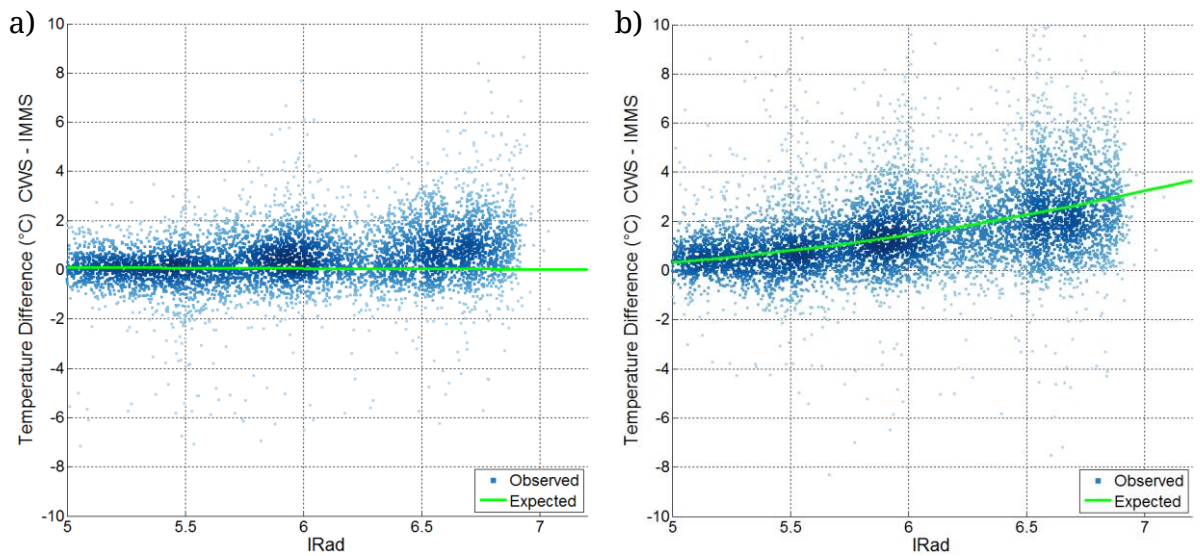


Figure 5.21. Relationship between the temperature difference (uncorrected CWS – IMMS) and l_{Rad} for stations whose metadata assigns them to the design class a) *Quality Louvered* and b) *Encased Louvered*. The green line represents the temperature bias we would expect to for this design type, learnt from the intercomparison field study (Figure 5.18).

5.5. Addressing representativity

Users of CWS data must consider representativity, asking themselves the question: ‘Are the observations a fair representation of the weather taking place over the region my application is trying to resolve’. This region may be a 1.5 km square grid cell in a high resolution numerical weather prediction model or a city district for which the urban heat island effect is being quantified. Crucially, the representativity of a CWS observation depends on the application in which it is used.

In Section 2.3.5 we detailed why representativity is so important and what makes it a difficult feature to quantify. Here in Section 5.5.1 we introduce a web application, named the *Station Classifier*, developed to allow users to classify a station’s exposure and Urban Climate Zone (UCZ); attributes that could act as proxies for representativity. In Section 5.5.2 the classifications made in this web application are used in an exploratory manner to investigate whether stations with different exposures and UCZs exhibit different bias signals with respect to IMMS. Finally we introduce how representativity error is actually modelled within this project.

5.5.1. Station classifier web application

The exposure and the Urban Climate Zone (UCZ) of a CWS are likely to play a large role in determining how representative the station is. Fortunately, as detailed in Section 2.1.4, WOW users can rank their station based on these two attributes. The exposure classes they can choose from are as follows:

Exposure

5: Very open exposure: no obstructions within $10h$ or more of temperature or rainfall instruments.

4: Open exposure: most obstructions/heated buildings $5h$ or from temperature or rainfall instruments, none within $2h$.

3: Standard exposure: no significant obstructions or heated buildings within $2h$ of temperature or rainfall instruments.

2: Restricted exposure: most obstructions/heated buildings $>2h$ from temperature or rainfall instruments, none within $1h$.

1: Sheltered exposure: significant obstructions or heated buildings within $1h$ of temperature or rainfall instruments.

0: Very sheltered exposure: site obstructions or sensor exposure severely limit exposure to sunshine, wind, rainfall.

R: Rooftop site: Rooftop sites for temperature and rainfall sensors should be avoided where possible.

T: Traffic site: equipment sited adjacent to public highway.

U: Exposure unknown or not stated.

h stands for the height of the obstruction above the sensor height. For example a 5 m tall obstruction, such as a tree, would need to be at least 7 m $((5 - 1.5) \times 2)$ from a 1.5 m tall CWS to be $>2h$ away. UCZ classes are defined as follows:

Urban Climate Zone (UCZ)

1: Intensely developed urban zone with detached close-set high-rise buildings with cladding, e.g. downtown towers.

2: Intensely developed high density urban with 2 - 5 storey, attached or very close-set buildings often of brick or stone, e.g. old city core.

3: Highly developed, medium density urban with row or detached but close-set houses, stores & apartments e.g. urban housing

4: Highly developed, low density urban with large low buildings & paved parking, e.g. shopping mall, warehouses.

5: Medium development, low density suburban with 1 or 2 storey houses, e.g. suburban housing.

6: Mixed use with large buildings in open landscape, e.g. institutions such as a hospital, university, airport.

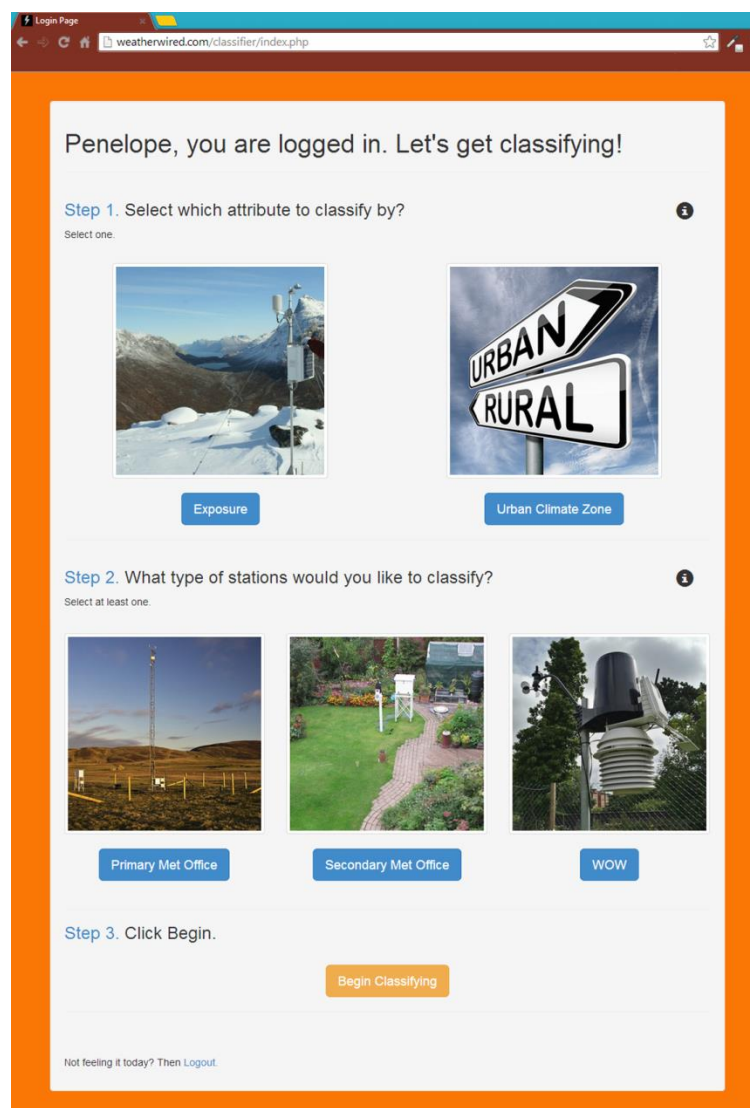
7: Semi-rural development with scattered houses in natural or agricultural area, e.g. farms, estates.

U: UCZ unknown or not stated.

Unfortunately around 15% of users failed to complete this optional ranking stage when setting up their weather station. The *Station Classifier* web application detailed here was designed to help overcome these data gaps. It allows a user to classify weather stations by exposure and UCZ using an aerial image of each site. The application can be used to classify both MMS and CWS sites providing a consistent set of classifications across all station types. This is particularly useful as operationally MMS sites are classified using a different scheme to WOW as specified by the World Meteorological Organisation (WMO, 2010).

To register and use the application yourself please follow the link below:

<http://www.weatherwired.com/classifier/>



The screenshot shows the 'Station Classifier' web application interface. At the top, a message says 'Penelope, you are logged in. Let's get classifying!'. Below this, there are three steps:

- Step 1. Select which attribute to classify by?**
Select one.
Two options are shown: 'Exposure' (with an image of a weather station on a snowy mountain) and 'Urban Climate Zone' (with an image of a road sign showing 'URBAN' and 'RURAL' directions).
- Step 2. What type of stations would you like to classify?**
Select at least one.
Three options are shown: 'Primary Met Office' (with an image of a tall weather station on a hill), 'Secondary Met Office' (with an image of a weather station in a garden), and 'WOW' (with an image of a weather station on a pole).
- Step 3. Click Begin.**
A single orange button labeled 'Begin Classifying' is present.

At the bottom, there is a link: 'Not feeling it today? Then Logout.'

Figure 5.22. Screenshot of the options page for the Station Classifier web app.

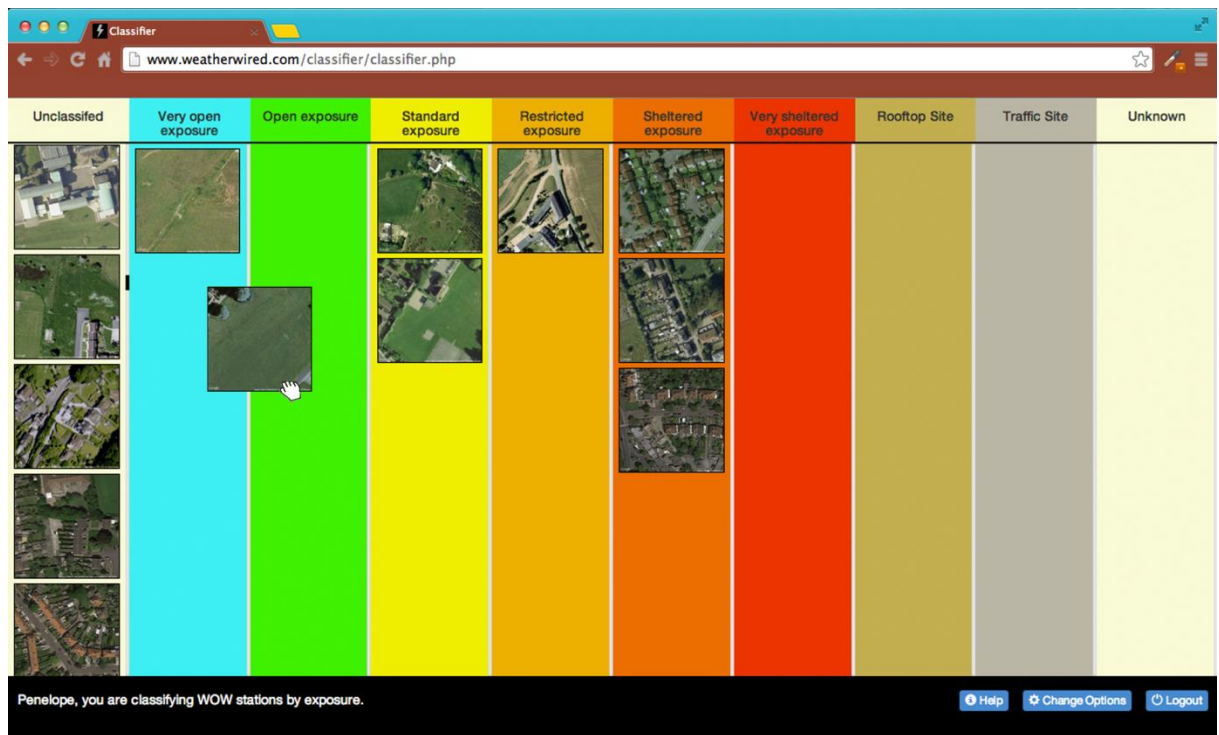


Figure 5.23. Screenshot of the Station Classifier web application being used to classify WOW stations by exposure. Note how the user is dragging an aerial image to the column they feel is most appropriate.

Figure 5.22 and Figure 5.23 illustrate the basic functionality of the web application. A user selects which station networks to classify and by which attribute, before being presented with aerial images of all the selected stations. It is then their job to look at the images, which can be enlarged or viewed in Google Maps, to assess which class the station belongs to, before dragging the image into the appropriate columns. The selling point of the application is the speed at which stations can be assessed, compared, and ultimately classified.

Although the application has currently only been trailed by a few team members, it is ready to be made publicly available. The eagerness of citizens to help in applications such as this has already been demonstrated by Fritz, et al., (2009) with their Geo-Wiki application, which enabled volunteers to validate and correct land cover maps using aerial images from Google Earth. By engaging the public, each station can be classified by multiple users allowing use to enjoy the ‘wisdom of the crowd’. This is important as the application is inherently subjective. For example, it is easy to assume that the marker in the centre of each image, which marks the station’s coordinates as provided in the metadata, is correct. For WOW stations these coordinates are specified by the station owner and may accidentally or deliberately include a degree of error. It is therefore up to the application user to question whether the coordinates are correct or if they believe the station actually resides nearby. To aid the decision the textual

metadata (Section 2.1.4) added by a WOW user is also visible. This can be useful, for example, for highlighting if a CWS is described as roof mounted, which can drastically change a site's exposure.

5.5.2. Exploratory analysis

Every station, professional and citizen alike, was classified using the *Station Classifier* application by a single user. Figure 5.24 shows their classifications for the MMS stations, and Figure 5.25 for the CWS. Unsurprisingly, given the strict WMO standards MMS stations adhere to, the majority of MMS stations were classified as well exposed, and located in *semi-rural* or *mixed-use* landscapes. The *mixed-use* class is used for stations located at airfields, of which there are numerous in the MMS network. CWS stations on the other hand were frequently classified with *restricted* or *sheltered* exposures in *suburban* environments, indicative of many owners' gardens. It is important to note that these classifications reflect the subjective opinion of the single user who completed the application. Ideally more users would have completed the application with any agreement between users helping to improve our confidence in the classifications. Even so it is clear that many of the possible categories are rarely used.

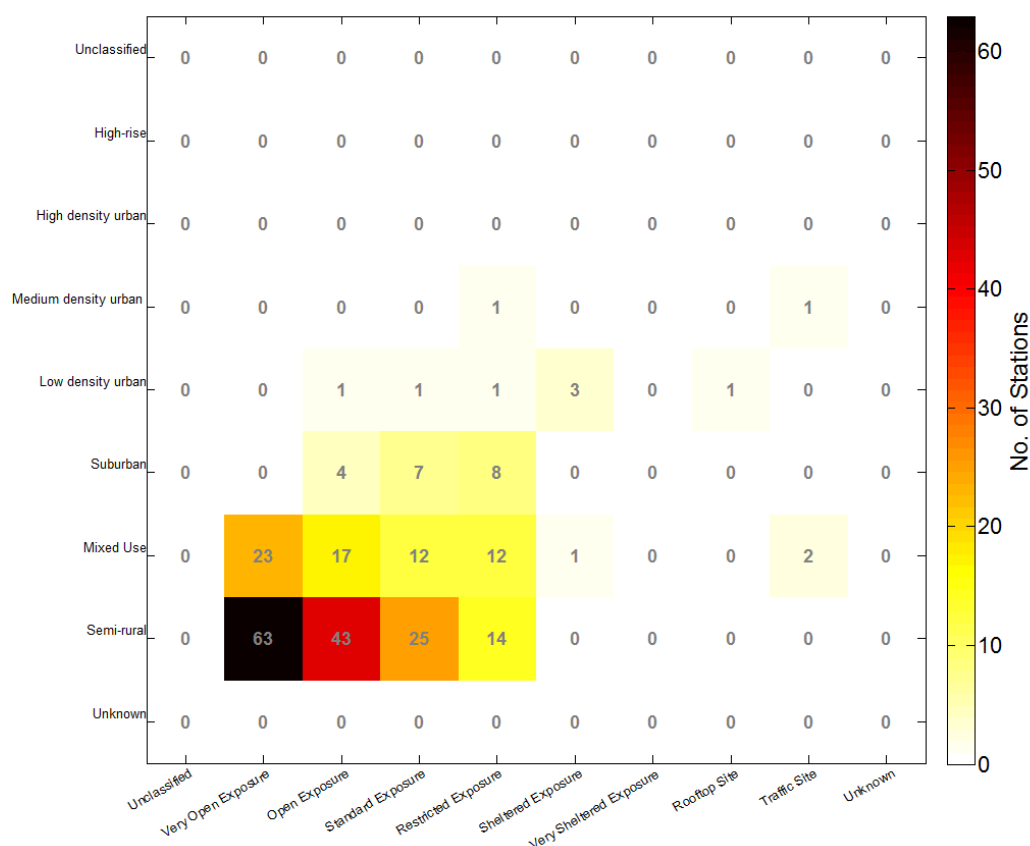


Figure 5.24. Distribution of MMS stations into the various exposure (columns) and Urban Climate Zone (rows) classes. The values represent the number of stations, with darker colours indicating a higher count.

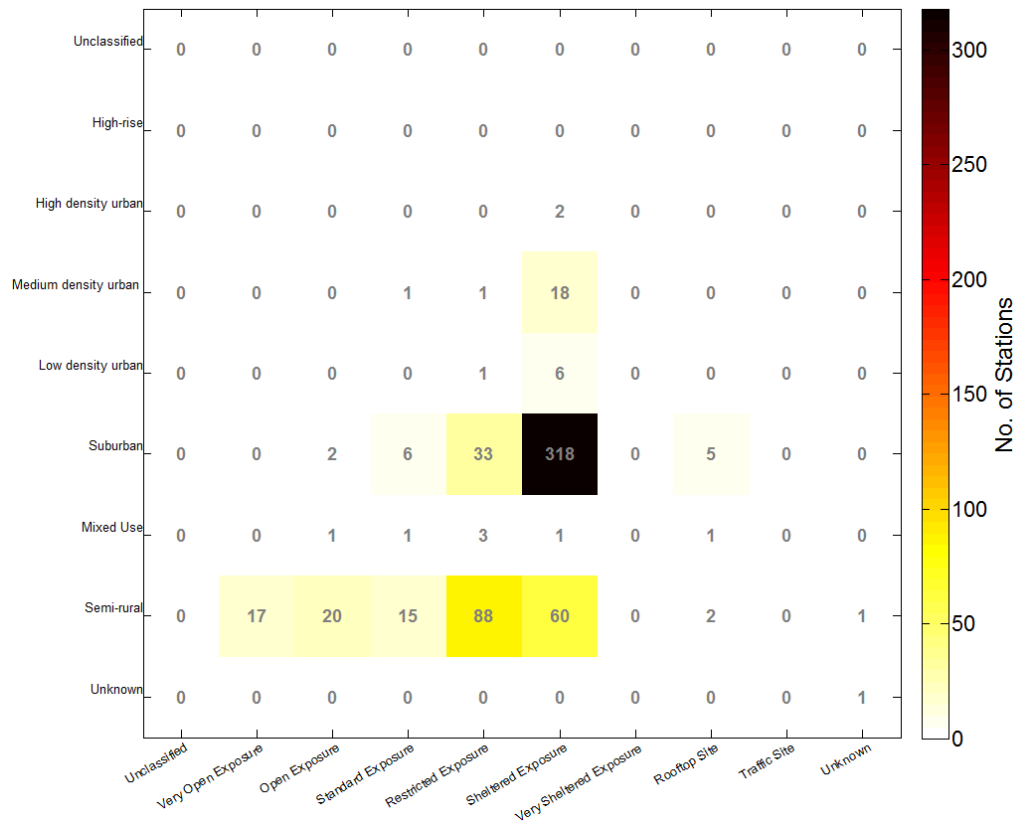


Figure 5.25. Distribution of CWS stations into the various exposure (columns) and Urban Climate Zone (rows) classes. The values represent the number of stations, with darker colours indicating a higher count. Every station counted submitted at least one observation during the summer period.

The following figures investigate whether the discrepancy between the CWS observations and IMMS is influenced by the stations' exposure and UCZ. This discrepancy can include instrumental errors, interpolation model errors and also the representativity errors of interest. Figure 5.26 and Figure 5.27 show that when the uncorrelated CWS data is compared against IMMS the instrumental errors, e.g. radiation-biases, dominate each box plot making it difficult to assess the variation between exposure and UCZ classes. It is therefore important that we try to remove these instrumental effects before making a comparison. In Figure 5.28 and Figure 5.29 the bias correction model, detailed later in Section 5.6, has been used to help remove these instrumental biases.

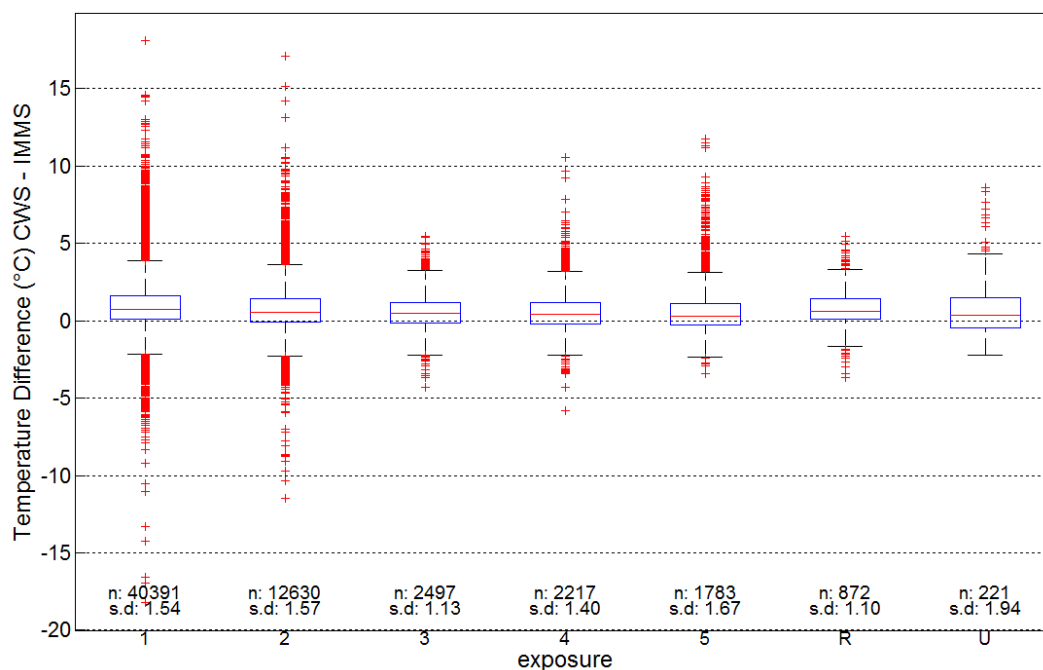


Figure 5.26. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their exposure classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown. The horizontal red line mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red.

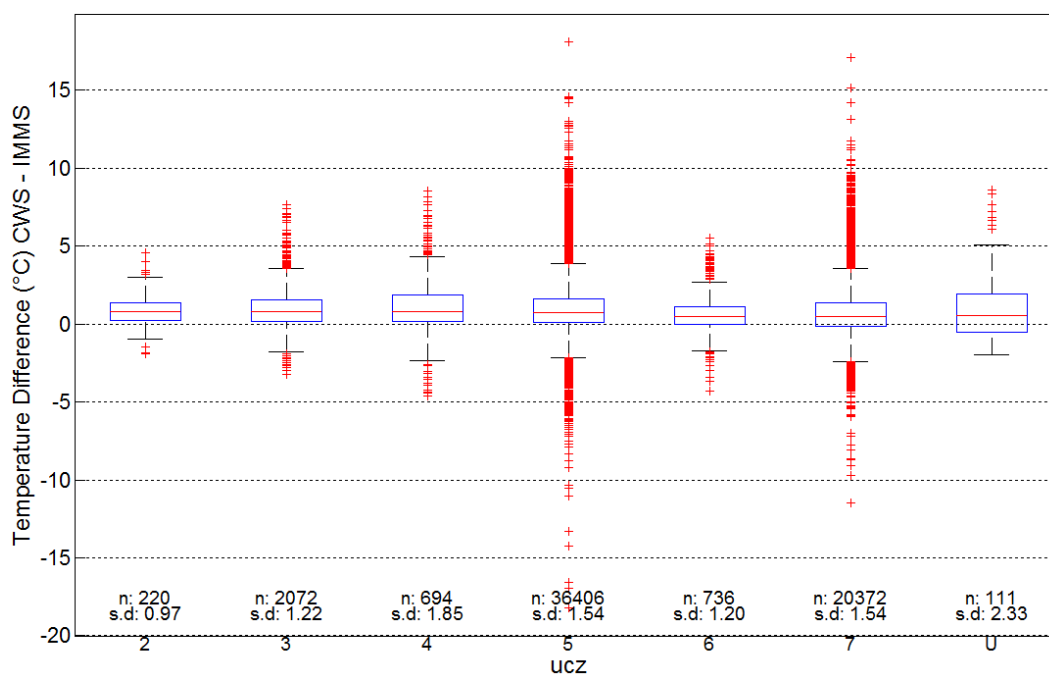


Figure 5.27. Boxplots of the temperature difference between uncorrected CWS observations and IMMS after the CWS stations have been separated by their UCZ classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.

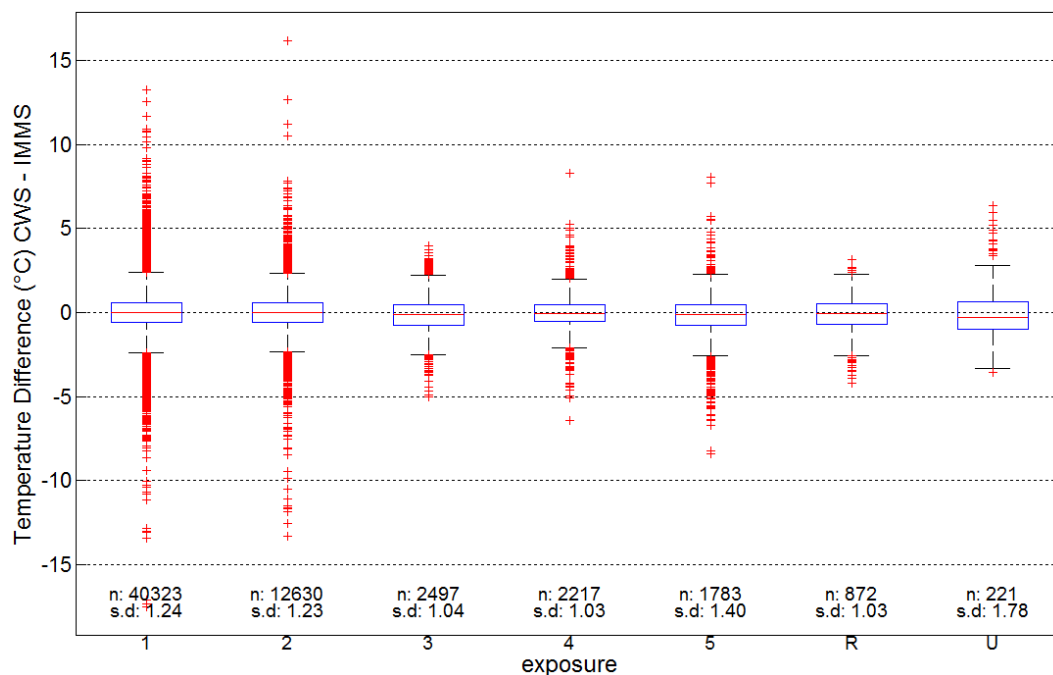


Figure 5.28. Boxplots of the temperature difference between corrected CWS observations and IMMS after the CWS stations have been separated by their exposure classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.

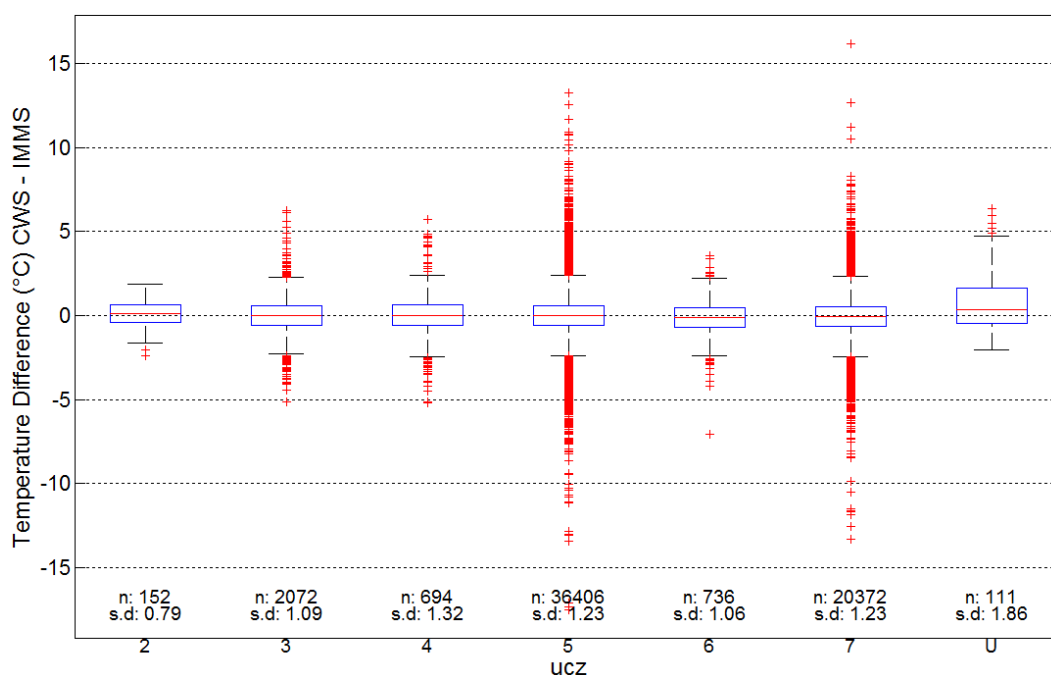


Figure 5.29. Boxplots of the temperature difference between corrected CWS observations and IMMS after the CWS stations have been separated by their UCZ classification as assigned within the Station Classifier app. The values below each box plot denote the number of observations, and the standard deviation, for each class. Only data from the summer period is shown.

Notice now that with the bias corrections applied, each class displays a median line close to 0 with minimal difference in their interquartile ranges. However, the *Unknown* class is an exception to this statement. There is a difference in the outliers (red points) between classes; however, this is largely a result of significant differences in the number of stations within each class. If instead we saw that certain classes displayed a biased median and/or an increased spread then we could increase our uncertainty about how representative stations within those classes are. As it is, these exploratory results suggest that these classifications provide little information about the representativity error and therefore the model must quantify representativity using the data itself. The following model framework section explains how an explicit representativity term is used to characterise the representativity error; using the difference between the CWS observations and IMMS to learn and update it.

5.6. Model framework

This section details the concept behind, and structure of, the complete bias correction model which uses as inputs: the interpolated MMS values (IMMS); radiation estimates; and the CWS observations themselves to learn temperature biases inherent to each CWS. The learnt biases can then be used to correct the CWS data such that the models' final output is a set of corrected CWS observations with associated uncertainty estimates. We begin by introducing the concept behind the model (Section 5.6.1) upon which the model's structure was built. Crucially it also highlights some key assumptions that have been made. Next the initial quality control procedure is detailed, essential for removing gross errors (Section 5.6.2). Section 5.6.3 defines the mathematical core of the model detailing the steps performed at each timestep in which the arrival of new data is used to update our estimate of each CWS's bias characteristics.

5.6.1. Concept

From the intercomparison field study, Section 3.2, it was clear that biases in the CWS temperature observations primarily result from either a calibration bias or a radiation-induced bias. As such our model aims to estimate these two biases explicitly and separately from one another. Each bias is represented by a Gaussian distribution with mean, μ , and variance, v , terms that are gradually learnt in a Bayesian manner from the CWS data itself. Figure 5.30 shows an example of Gaussian distributions that may have been learnt by the bias correction model. The means of the two biases are subtracted from the uncorrected CWS observation to produce the corrected estimate. The variances are added. Note that the radiation bias mean is non-zero, as the majority of CWS displayed positive radiation-induced biases. The uncorrected CWS

observation is initially given a variance based on instrumental sensor noise, i.e. of the thermistor itself.

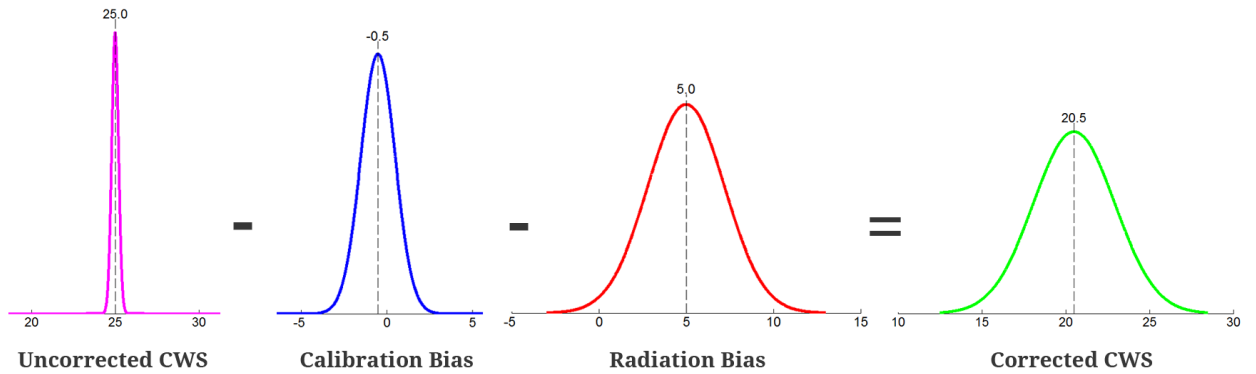


Figure 5.30. Example of the learnt biases, modelled as Gaussian distributions, being subtracted from the raw CWS observation in order to correct it.

For the calibration bias we assume *a priori* that every station new to the model displays a zero mean bias. Over time the data can update this belief, changing it from a zero mean if the data suggests as much. During the day stations with poor radiation shielding usually experience radiation-induced biases well in excess of any calibration bias. As such, the mean of calibration bias (referred to below as $\mu_{s,t}^{Cal}$) is only learnt during the night so that any radiation-induced biases are not misinterpreted by the model as a warm calibration bias. This is based upon the assumption that the calibration bias is stable through time; i.e. the value learnt during the night fairly represents the calibration bias experienced during the day.

Unlike the mean, the variance term of the calibration bias (referred to below as $v_{s,t}^{Cal}$) is updated during the day as well. This variance term is used as a proxy for representativity error; used to explicitly quantify our uncertainty using the data itself. The assumption is that, having explicitly learnt the calibration and radiation-induced bias, this term mops up any discrepancies that remain, with representativity error the most probable cause. It is crucial to note however that it is a measure of the representativity error *relative to* the IMMS predictions from the temperature interpolation model. Therefore any natural spatial variations that the interpolation model is not competent to resolve, as well as any errors in the interpolation model, are mopped up by this representativity term. As we expect the representativity term to be less stable through time than the calibration mean, perhaps even exhibiting a diurnal pattern, this is updated using both day and night observations.

Unlike the calibration bias the radiation-induced temperature bias is not updated directly. Instead we update the probability that a given station belongs to each of the

shielding design classes listed in Section 5.4.2. For a given radiation estimate, l_{Rad} , each design class predicts a different magnitude of bias. Once the calibration bias has been removed, the discrepancy between the CWS observation and IMMS provides a current estimate the radiation bias, albeit with a degree of uncertainty. For design classes that predict a bias similar to the current estimate their probability is increased. Gradually these membership probabilities are adjusted influencing the impact each design class has on the estimate of the radiation bias. Section 5.6.3 explains exactly how this is done. The membership probabilities are only updated during significant levels of incoming radiation. Under low radiation conditions, the difference in bias between design classes is largely indistinguishable, making it difficult to assign a CWS to the correct class. By learning the radiation bias indirectly through design class probabilities a series of assumptions are made. Firstly the characteristics of each design class are fixed, although this ensures that the radiation-induced biases we predict are based on empirical evidence the assumption is thus made that the biases experienced by real CWS closely fit those which characterise one of these predesigned design classes. Following on from this the model therefore assumes that the radiation-induced biases experienced by real CWS does not exceed the range experienced during the intercomparison field study on which the design classes are based. As interesting topic for further work would be to develop a system in which the characteristics of each design class could evolve through time if there is evidence from the data that they should. This would allow bias characteristics which were not identified during the field study to be corrected for. The significant risk with this approach, and why it wasn't favoured here, is that each design class may lose its original identity and physical meaning.

Whenever we refer to the calibration bias and design probabilities being 'updated' a Bayesian framework is used. The prior information is usually the posterior from the last timestep forecast forward to the current timestep. It is important that as we forecast forward in time our uncertainty grows as implemented in the forecast steps of Section 5.6.3. By only combining preceding information with our current estimate the model acts as a filter rather than smoother. Unlike a smoother a filter can be used operationally in real-time. To control the rate at which the priors are forgotten, and updated with current estimates, forgetting rate and learning rate parameters, denoted as γ and α respectively are used. These are set differently for the calibration bias and design class updates. This ensures that the calibration bias, thought to be more stable, updates more gradually. The inclusion of γ and α is based on the assumption that the model contains correlated errors. The timestep between observations (referred to below as δt) is used alongside γ and α to ensure the rate of update is dependent on

elapsed time and not the number of observations. Without it the learnt biases for stations with a high upload frequency (e.g. every minute) would update too quickly with the risk of interpreting a natural short-lived signal as calibration bias. It also enables the model to sensibly handle multiple stations uploading at different frequencies, data which often also includes missing data. If there is no data at a given timestep then δt is simply increased.

To illustrate how this concept would work in practice, Figure 5.31 shows an idealistic representation of how the model would react when presented with biased data. It shows how the calibration bias and radiation bias are learnt separately over time, so that by the end of the period a reliable correction is being made.

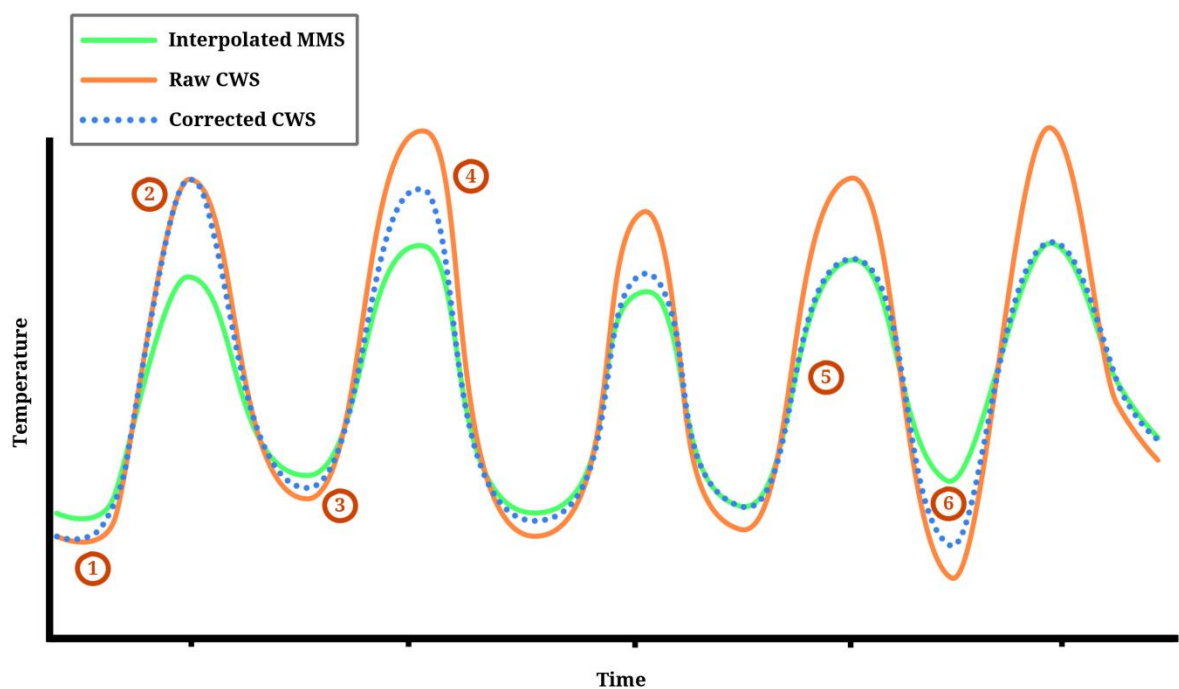


Figure 5.31. Theoretical time series over 5 days showing the CWS observation before and after a correction has been applied. The numbers correspond to the numbered steps below, which detail what is occurring at each step.

The following points correspond to those labelled on the figure above and traverse the basic model logic. Note that in reality the speed at which these biases are updated would tend to be slower.

- 1) The model is presented with CWS data with a consistent negative calibration bias. Thus the CWS data appears colder than the interpolated MMS values, although during the day the radiation bias hides this effect. Initially the prior for the calibration bias was set as 0 so no correction is made.

- 2) The CWS data also contains a strong radiation bias causing a warm bias during the day. As with the calibration bias, the radiation bias was initially assumed to be 0, so at first no correction is applied.
- 3) After some time the model is beginning to learn that there is a calibration bias which it starts to correct for.
- 4) Likewise the model also begins to learn and correct the radiation bias by adjusting what station design it believes the CWS is, and in turn applying the corresponding correction.
- 5) By this point the calibration and radiation biases have been reliably learnt so that after the corrections have been applied the CWS data resembles that of the interpolated MMS data.
- 6) After the corrections have been applied there is still a discrepancy between the IMMS and the CWS data at this time. Assuming a reliable correction has been applied then this difference must be a natural spatial variation that we are trying to capture. A key assumption here is that the learnt calibration and radiation biases are more stable through time than local natural spatial variations; i.e. a natural spatial variation comes and goes before the calibration and radiation biases have time to adjust significantly to this discrepancy between IMMS and the CWS observation.

In the model performance section (Section 5.7.1) it is possible to see signs of this occurring with real CWS data.

5.6.2. Initial CWS quality control

CWS data can contain gross errors. Many of the probable causes of such errors were highlighted in Section 2.3. To remove these errors a set of simple quality control checks are used to pre-process the CWS data before it enters the bias correction model. The goal here is not to remove any biases that can be parameterised or learnt over time, but to only remove unrealistic gross errors that if left in would be falsely interpreted by the model as a calibration or radiation-induced bias. For checks that compare the CWS data against IMMS values, we were careful to set thresholds that could accommodate for the natural spatial variations which should not be excluded.

The quality control checks used are as follows:

- Upper and Lower Bounds – CWS data that lie outside of sensible bounds are removed. These upper and lower limits were set by inflating the temperature range displayed by the MMS data at each timestep. Here the lower limit was set as 5 °C below the minimum MMS value, and the upper limit as 12 °C above the

maximum MMS value. The choice of these limits was somewhat ad hoc, based roughly upon the magnitude of the biases seen in the intercomparison field study. When setting the bounds it is important to consider that the timing of the CWS observations may differ slightly from that of the professional observations.

- Persistence – A station that submits the same observation more than n times consecutively has these consecutive readings removed. n is set generously and should be based up on the temporal resolution of observations. Here hourly CWS data underwent this check, setting $n = 6$, before filtering down to the final 3 hourly resolution.
- Spikes – Step changes in temperature greater than a specified tolerance are assumed to represent artificial spikes and are therefore removed. Again this tolerance should consider the temporal resolution of the CWS data. Here a step change of 10 °C per hour was assumed to be erroneous.
- Majority Missing – For stations with a lot of missing data (i.e. over 97%) the assumption is made that the data which is uploaded is prone to gross errors and thus all of that station's data is discounted.
- Correlation – A weak correlation between the CWS data and IMMS proved to be a good indicator of gross errors. In particular it helps highlight timing issues such as observations submitted with an incorrect timestamp, perhaps due to a difference in timezone. When the correlation coefficient falls below a given threshold (set here as 0.4) the data is excluded. This check is performed over a moving window so that the whole dataset need not be deleted when the data only shows a poor correlation temporarily. The length of this window should consider the temporal resolution of the observations, for example if both CWS and professional data are available at minute intervals then the window may be much shorter than if they submitted at hourly intervals.
- Mean vs Modal – When the mean timestep between observations is dramatically different from the mode this is indicative of sporadic data submissions, which are often prone to gross errors. Again, this should be performed over a moving window, which can be shorter when the temporal resolution is higher.

Any operational implementation of our complete quality control system should include gross error checks such as these. Many of the checks can be performed in real-time as soon as a new observation arrives; however some checks such as *Correlation* and *Mean vs Modal* have a lagged effect as they rely on a longer time series to be effective.

As previously described (Section 2.1.4) citizen observers are encouraged to share the elevation of their station when they subscribe to WOW, however as was evident in Section 2.3.4, not all citizens provide an elevation value, and some of those that do appear to do so incorrectly. Therefore a quality control check is required to correct these WOW elevations. A very simple approach was used, in which WOW metadata elevations were compared against the DEM height (see Section 4.4.2) at their location. If the metadata height was out by more than 50m then the DEM was used instead. Figure 4.12 confirmed that these DEM heights are reliable. With an accurate elevation the temperature interpolation model can better estimate IMMS at these CWS locations leading to better estimates of the calibration and radiation-induced biases.

5.6.3. Bayesian update procedure

At each timestep the bias correction model performs a series of steps in order to update and predict the calibration and radiation-induced temperature biases producing a final set of corrected CWS observations with uncertainty estimates. Here is a short summary in chronological order of the steps performed followed by a detailed explanation of the mathematical operations performed during each step.

1. **Interpolation** - The Bayesian linear regression model is run twice, once to interpolate MMS temperature observations, and secondly to interpolate MMS radiation observations, both to the locations of the citizen stations.
2. **Predict Radiation Bias** – Using the design probabilities learnt from previous timesteps the radiation-induced bias is predicted for each CWS station.
3. **Forecast Calibration Bias** – Given the arrival of new data it is possible to establish how long it is been since the last observation. This timestep is used to inflate the uncertainty (representativity) term as the learnt bias from the previous timestep is propagated forward to act as the prior at this timestep.
4. **Update Calibration Bias** – The predicted radiation bias is removed from the new CWS data providing a current estimate of the calibration bias. This is used to update the prior estimate. Note that the update of the mean term is only performed at night.
5. **Forecast Design Probabilities** – As the design probabilities from the previous timestep are brought forward to the current timestep the model becomes more uncertain about which design class each station belongs to.
6. **Update Design Probabilities** – Having removed the model estimate of the calibration bias the design probabilities can now be updated for each station.
7. **Re-Predict Radiation Bias** – The updated design probabilities are then used to produce an up-to-date estimate of the radiation bias.

8. **Correct CWS observation** – The mean estimates of the calibration and radiation biases are subtracted from the CWS observation producing a corrected value. The variances are added to produce a final variance indicative of the overall uncertainty.

The equations used in the following steps adhere to the notation outlined in Appendix 8.1, please refer to it as required. Figure 5.32 should also be referenced to as it illustrates the flow of data through the following functions.

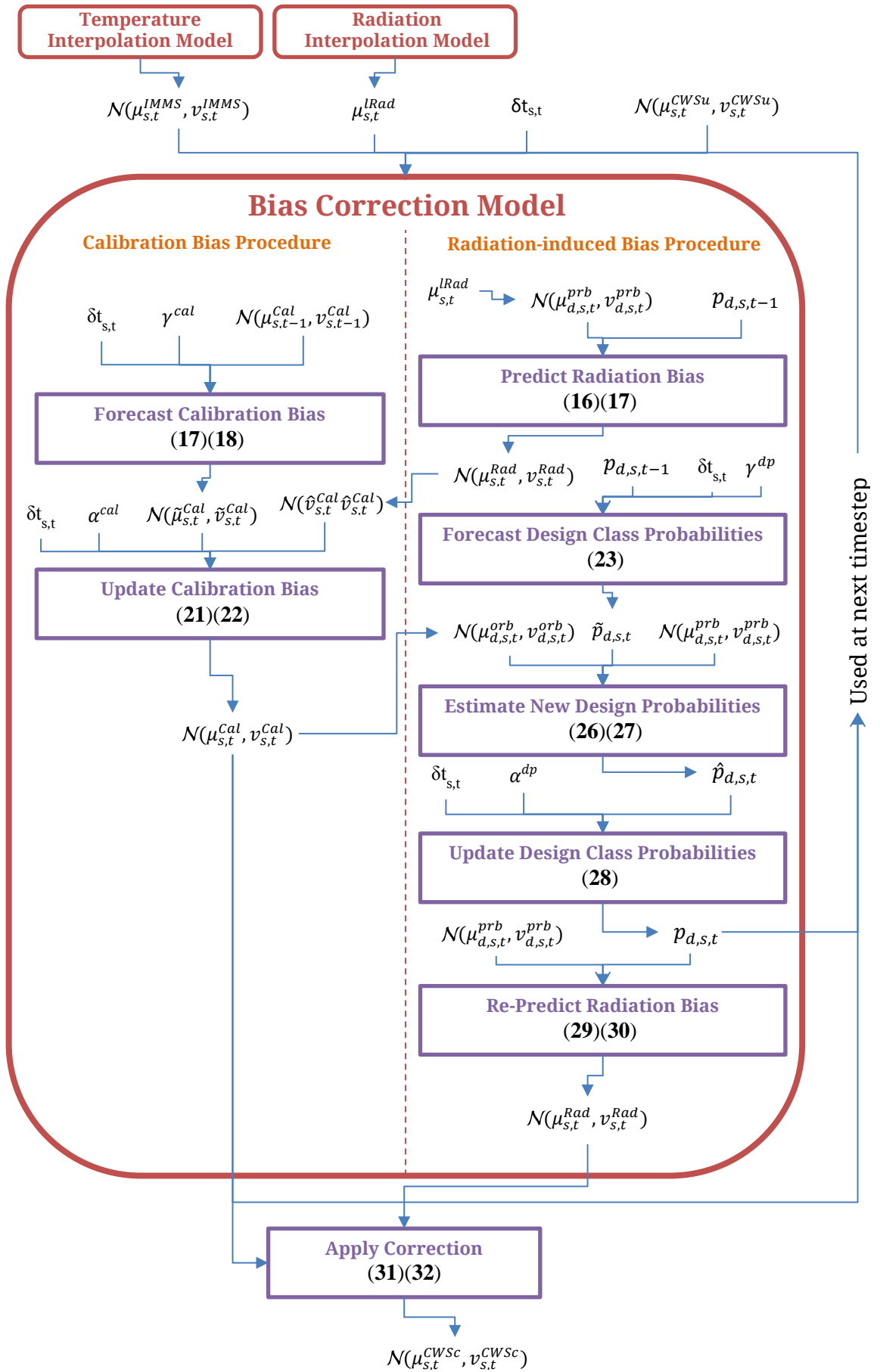


Figure 5.32. Schematic of data flow through the bias correction model. Bracketed numbers refer to equation numbers in Section 5.6.3.

1. Interpolation

The Bayesian linear regression model is first run as detailed in Chapter 4 to produce an independent estimate of the temperature at CWS locations by interpolating MMS observations (i.e. IMMS). This estimate is a Gaussian with mean, $\mu_{s,t}^{IMMS}$, and variance $v_{s,t}^{IMMS}$.

The interpolation model is run a second time, as detailed in Section 5.3, to provide l_{Rad} estimates at the CWS locations. These l_{Rad} estimates are Gaussian distributions, however only the mean is used. Configuring the bias correction model so that it also accounts for the uncertainty is the subject for further work.

2. Predict Radiation Bias

The design probabilities learnt at the previous timestep are used to make an initial estimate of the radiation-induced temperature bias with mean, $\mu_{s,t}^{Rad}$, and variance, $v_{s,t}^{Rad}$. This estimate is required in order to estimate the calibration bias in equations (19) and (20). It is computed using:

$$\mu_{s,t}^{Rad} = \sum_d \left(\mu_{d,s,t}^{prb} \cdot p_{d,s,t-1} \right), \quad (15)$$

$$v_{s,t}^{Rad} = \sum_d \left(v_{d,s,t}^{prb} \cdot p_{d,s,t-1} \right) + \sum_d \left(p_{d,s,t-1} \left(\mu_{s,t}^{Rad} - \mu_{d,s,t}^{prb} \right)^2 \right), \quad (16)$$

where $p_{d,s,t-1}$ represents the probability that a given station, s , belongs to a design class, d , as learnt at the previous timestep, $t-1$. prb stands for the Predicted Radiation Bias. It is the radiation-induced temperature bias predicted by each of the design classes using the estimate of l_{Rad} at each CWS location. These predictions have a mean, $\mu_{s,t}^{prb}$, and a variance, $v_{s,t}^{prb}$. The latter accounts for the uncertainties quantified during the field study from which the relationship between l_{Rad} and the temperature bias for each design class are based. The equations used to make these predictions have the same form as Equations (10) and (11) from the predict step in the interpolation model.

3. Forecast Calibration Bias

The following equations ensure that as the model moves forward in time it becomes more uncertain about the exact value of the calibration bias – that is until the arrival of new data in Step 4.

The prior mean, $\tilde{\mu}_{s,t}^{Cal}$, is simply the same as the posterior from the previous timestep:

$$\tilde{\mu}_{s,t}^{Cal} = \mu_{s,t-1}^{Cal} . \quad (17)$$

The prior variance, $\tilde{v}_{s,t}^{Cal}$, is calculated as follows:

$$\tilde{v}_{s,t}^{Cal} = v_{s,t-1}^{Cal} + v_{s,t=0}^{Cal} \frac{\delta t_{s,t}}{\gamma^{Cal}} , \quad (18)$$

where δt is a vector containing the time since each CWS last had an observation. γ^{Cal} is the forgetting rate parameter specific to the calibration bias and is constant. Here it is set as 50 days. Essentially the longer it has been since the model last saw an observation for a given station the more uncertain it becomes about that station's calibration bias, with the rate at which it becomes more uncertain specified by γ^{Cal} . $v_{s,t=0}^{Cal}$ is the calibration bias variance as set at the initial timestep, and is used to scale the degree to which the uncertainty grows.

4. Update Calibration Bias

Updating the model calibration bias requires a current estimate of the ‘observed calibration’ bias at this timestep, $\mathcal{N}(\hat{\mu}_{s,t}^{Cal}, \hat{v}_{s,t}^{Cal})$. The mean, $\hat{\mu}_{s,t}^{Cal}$, is estimated using the discrepancy between the uncorrected CWS observation, $\mu_{s,t}^{CWSu}$, and IMMS, $\mu_{s,t}^{IMMS}$, whilst accounting for the estimated radiation bias, $\mu_{s,t}^{Rad}$, as follows:

$$\hat{\mu}_{s,t}^{Cal} = \mu_{s,t}^{CWSu} - \mu_{s,t}^{IMMS} - \mu_{s,t}^{Rad} . \quad (19)$$

The uncertainties from each are added.

$$\hat{v}_{s,t}^{Cal} = v_{s,t}^{CWSu} + v_{s,t}^{IMMS} + v_{s,t}^{Rad} . \quad (20)$$

Equations (21) and (22) show how these current estimates are combined with the prior to give the posterior estimate of the calibration bias.

$$\mu_{s,t}^{Cal} = \tilde{\mu}_{s,t}^{Cal} + (\hat{\mu}_{s,t}^{Cal} - \tilde{\mu}_{s,t}^{Cal}) \frac{\tilde{v}_{s,t}^{Cal}}{\tilde{v}_{s,t}^{Cal} + \hat{v}_{s,t}^{Cal}} \frac{\delta t_{s,t}}{\alpha^{Cal}} \quad (21)$$

$$v_{s,t}^{Cal} = \left(1 - \left(\frac{\tilde{v}_{s,t}^{Cal}}{\tilde{v}_{s,t}^{Cal} + \hat{v}_{s,t}^{Cal}} \frac{\delta t_{s,t}}{\alpha^{Cal}} \right) \right) \tilde{v}_{s,t}^{Cal} \quad (22)$$

α^{cal} is the learning rate parameter specific to the calibration bias, set here as 24 hours. It dictates the rate at which calibration bias can update. $\tilde{\mu}_{s,t}^{cal}$ and $\tilde{v}_{s,t}^{cal}$ are obtained from Equations (17) and (18).

Forecast and update steps are only performed for stations which have observations at the current timestep. This is because δt is only known when a new observation arrives and an update can only be made when new data is available. As explained previously the mean, $\mu_{s,t}^{cal}$, is only updated at night, whereas the variance, $v_{s,t}^{cal}$, updates at all times. It should be noted that these equations are in essence the classical Kalman filter update equations, with a learning rate parameter.

5. Forecast Design probabilities

Equation (23) shows that as the model moves forward in time it becomes more uncertain about which design classes are most probable. It does so by evening out their individual probabilities. This ensures the model does not become overconfident about the most probable design class:

$$\tilde{p}_{d,s,t} = \left(1 - \frac{\delta t_{s,t}}{\gamma^{dp}}\right) p_{d,s,t-1} + \frac{\delta t_{s,t}}{\gamma^{dp}} p_d^{eq}, \quad (23)$$

where p_d^{eq} denotes the probability of belonging to each design class when each is equally probable, i.e. when the model has learnt nothing. γ^{dp} is the forgetting rate parameter specific to the design probabilities. Here it is set as 20 days. It dictates the rate at which the probabilities return to being equal, i.e. to p_d^{eq} . To ensure the consistency of the model $\frac{\delta t_{s,t}}{\gamma^{dp}}$ is restricted to ≤ 1 . In practise $\frac{\delta t_{s,t}}{\gamma^{dp}} \ll 1$ so the forecast step only makes small changes to $p_{d,s,t}$. This forecast step is essential, let us say for example that the model converged on a single design type for a given CWS, then without this step the model would never be able to update and change its belief later on, even if the citizen had modified their station's design type causing completing different bias characteristics.

6. Update Design Probabilities

In order to update the design probabilities, $p_{d,s,t}$ the model requires an estimate of the 'observed radiation bias' (orb) at the current timestep with a mean, $\mu_{s,t}^{orb}$, and variance, $v_{s,t}^{orb}$. It is calculated as the discrepancy between the uncorrected CWS observation and IMMS accounting for the calibration bias that has already been updated for this timestep as previously detailed in Equations (21) and (22):

$$\mu_{s,t}^{orb} = \mu_{s,t}^{CW^{Su}} - \mu_{s,t}^{IMMS} - \mu_{s,t}^{Cal}, \quad (24)$$

$$v_{s,t}^{orb} = v_{s,t}^{CW^{Su}} + v_{s,t}^{IMMS} + v_{s,t}^{Cal}. \quad (25)$$

The model must now quantify the degree to which the ‘observed radiation bias’ estimate, $\mathcal{N}(\mu_{s,t}^{orb}, v_{s,t}^{orb})$, agrees with predicted radiation bias for each individual design class, $\mathcal{N}(\mu_{d,s,t}^{prb}, v_{d,s,t}^{prb})$. The model assumes that the probability of belonging to each design class, given the current estimate, can be modelled with the scale factor S of the product of two Gaussian distributions (Bromiley (2013)). Figure 5.33 illustrates how the approach compares the product of the two Gaussians; weighting design classes with a larger shared area more heavily.

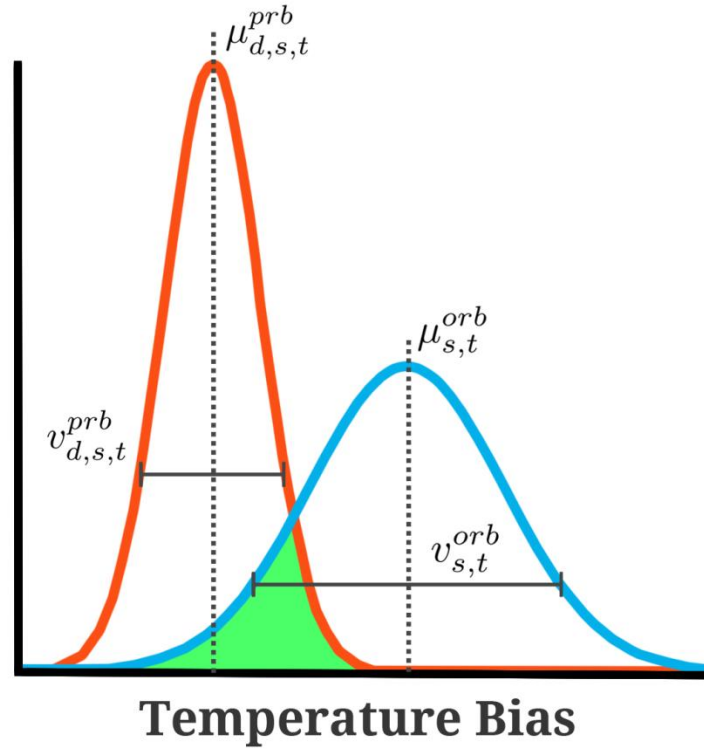


Figure 5.33. Visual representation of the overlap between the Predicted Radiation Bias for a given design class (Orange curve; e.g. *Encased*) and the estimated Observed Radiation Bias (Blue curve). The size of green overlap area is proportional to the scaling factor S.

The scaling factor S is calculated in practice as:

$$S_{d,s,t} = \exp \left(-\frac{1}{2} \frac{\left(\mu_{s,t}^{orb} - \mu_{d,s,t}^{prb} \right)^2}{v_{s,t}^{orb} - v_{d,s,t}^{prb}} \right). \quad (26)$$

It is important to normalise the scaling factors for each design class so that together they sum to one (Equation (27)). This provides the new set of design probabilities, $\hat{p}_{d,s,t}$, used to update the probabilities that have been learnt previously, $\tilde{p}_{d,s,t}$, (Equation (28)).

$$\hat{p}_{d,s,t} = S_{d,s,t} / \sum_d S_{d,s,t} \quad (27)$$

$$p_{d,s,t} = \left(1 - \frac{\delta t_{s,t}}{\alpha^{dp}}\right) \tilde{p}_{d,s,t} + \frac{\delta t_{s,t}}{\alpha^{dp}} (\tilde{p}_{d,s,t} \cdot \hat{p}_{d,s,t}) \quad (28)$$

α^{dp} is the learning rate parameter specific to the design probability update (set here as 6 hours) and acts in relation to the time since the last observation, δt , to control the rate at which the design probabilities can change. As with the forecast step $\frac{\delta t_{s,t}}{\alpha^{dp}}$ is restricted to ≤ 1 .

The multiplication of $\tilde{p}_{d,s,t}$ with $\hat{p}_{d,s,t}$ is important as it ensures that a CWS station tends to dominantly belong to just one design class for as long as current estimates agree with the prior. When this approach was tested with simulated data, a station which belonged strongly to a single class tended to give a better estimate of the radiation-induced bias than when it belonged to several.

As there would be little difference in $\mathcal{N}(\mu_{d,s,t}^{prb}, v_{d,s,t}^{prb})$ between design classes when l_{Rad} is small the forecast and update steps for the design probabilities are only performed when l_{Rad} is above a certain threshold. Here this threshold was set as $l_{Rad} > 5.5$, equivalent to $\sim 150 \text{ W m}^{-2}$.

7. Re-Predict Radiation Bias

Now that the design probabilities have been updated the radiation bias is re-predicted to produce an up-to-date estimate, $\mathcal{N}(\mu_{s,t}^{Rad}, v_{s,t}^{Rad})$. These equations are equivalent to Equations (15) and (16) except that $p_{d,s,t}$ replaces $p_{d,s,t-1}$:

$$\mu_{s,t}^{Rad} = \sum_d \left(\mu_{d,s,t}^{prb} \cdot p_{d,s,t} \right) \quad (29)$$

$$v_{s,t}^{Rad} = \sum_d \left(v_{d,s,t}^{prb} \cdot p_{d,s,t} \right) + \sum_d \left(p_{d,s,t} \left(\mu_{s,t}^{Rad} - \mu_{d,s,t}^{prb} \right)^2 \right) \quad (30)$$

8. Correction CWS observation

The model now has posterior distributions for both the calibration and radiation-induced biases. Subtracting these two biases from the uncorrected observation, $\mu_{s,t}^{CWSu}$, results in the final corrected CWS observation with a mean term $\mu_{s,t}^{CWSc}$:

$$\mu_{s,t}^{CWSc} = \mu_{s,t}^{CWSu} - \mu_{s,t}^{Cal} - \mu_{s,t}^{Rad} \quad (31)$$

The final uncertainty, $v_{s,t}^{CWSc}$, is a combination of $v_{s,t}^{CWSu}$, which represents CWS sensor noise (set as 0.2 °C), and the uncertainties of the calibration and radiation-induced bias estimates:

$$v_{s,t}^{CWSc} = v_{s,t}^{CWSu} + v_{s,t}^{Cal} + v_{s,t}^{Rad} \quad (32)$$

These ‘corrected’ CWS values are then able to be used in place of the raw CWS data, with the added benefit of including an estimate of the uncertainty in the corrected value.

5.6.4. Computational resources

Here we detail the computational resources we used to run this system. As our aim was simply to demonstrate this approach, rather than run it operationally, we did not run this system in real-time. Instead it was fed past data from our case study periods (Section 4.2).

In our implementation the following data was pre-processed ready for our model to use:

- MMS temperature observations (Section 4.1).
- MMS radiation observations (Section 5.3).
- Uncorrected CWS temperature observations (Section 5.1).
- Initial estimate of the station design class (Section 5.4)
- UKV model output (Section 4.4.1).
- Easting, Northing, Elevation, Coastality, Urbanisation and RBF estimates at every station location (Sections 4.4.2 – 4.4.5).
- Clear-Sky GHI estimates (Section 5.3.1).
- Visible and infrared satellite imagery (Section 5.3.2).

With this pre-processing performed, our system would then perform the following key processes at each timestep:

- Interpolate MMS temperature observations to CWS locations (Section 4).
- Interpolate MMS radiation observations to CWS locations (Section 5.3).
- Quantify and correct for bias in CWS observations and quantify associated uncertainty (Section 5.6).

A single 2 week case study period comprised of 111 timesteps for which these 3 key processes could be performed for all timesteps (in sequence) in under 5 seconds. At each timestep it corrected ~600 CWS stations, using ~240 MMS stations with temperature observations, and ~80 MMS stations with radiation observations. Therefore with an update time for a single iteration of less than 0.05 s this approach is entirely suitable for real-time use, this is thanks to choice of linear models which are sufficiently robust for such a use.

This was performed using MathWorks' MATLAB software as it provides an excellent interactive environment for writing, running, debugging and visualising mathematical models. It is also well suited to matrix operations as used frequently in this system. This was run on a Windows 7 desktop computer with a 3.3GHz quad-core Intel Core i5-2500 processor with 16GB of RAM.

Therefore, the potential bottleneck for running this system in real-time is not the bias correction and interpolation models themselves, but in retrieving and pre-processing the input data required. The speed of the latter would depend on the existing resources any organisation wishing to implement such an approach has available.

The following big O notation details how the computation speed and memory requirements would scale with increasing numbers of CWS stations, N_{CWS} , MMS stations, N_{MMS} , and basis functions, N_β :

Computational: $O(N_{CWS}) + O(N_{MMS}^2) + (N_\beta^3)$

Memory: $O(N_{CWS}) + O(N_{MMS}^2) + (N_\beta^2)$

5.7. Model performance

Here we use two approaches to assess the performance of the bias correction model. Firstly in Section 5.2 we saw obvious signs that the uncorrelated data exhibited biases with respect to IMMS, in particular many stations displayed daytime warm biases. Having corrected the CWS data we would hope to see that on average the CWS data is

now unbiased (Section 5.7.1). This first approach is not ideal as it compares the data against IMMS which itself is not a perfect estimate of the true temperature at the CWS location. Even if the bias correction model perfectly learnt the calibration and radiation-induced biases, we would still expect a discrepancy due to interpolation model errors and representativity errors, as well the crucial natural spatial variations that we wish to capture and that provide the added value.

The second approach, Section 5.7.2, aims to quantify the added value these CWS observations bring, and to show that only once the observations have been corrected do we see the real benefit of their inclusion. To do this, the CWS observations are fed back into the temperature interpolation model to see whether their inclusion improves the cross-validation error over just using the MMS stations alone. The input CWS data used is the same 3-hourly WOW data, over the four 2 week case periods, as detailed in Section 5.2.

5.7.1. Corrected CWS vs. Interpolated MMS

Here we compare the corrected CWS data against IMMS to verify that after the bias correction model's corrections have been applied the CWS data no longer exhibits any systematic biases. Figure 5.34 and Table 5 clearly show that before the CWS data was corrected it contained a systematic warm bias most notable during the higher temperatures experienced within the warmer periods, with radiation-induced biases the probable cause. Once our learnt corrections have been applied this systematic bias is successfully removed. The standard deviation of the discrepancy is also reduced. The assumption here is that once the learnt instrumental biases have been removed the natural spatial variations and model errors that make up the remaining discrepancies do not themselves contain any systematic biases.

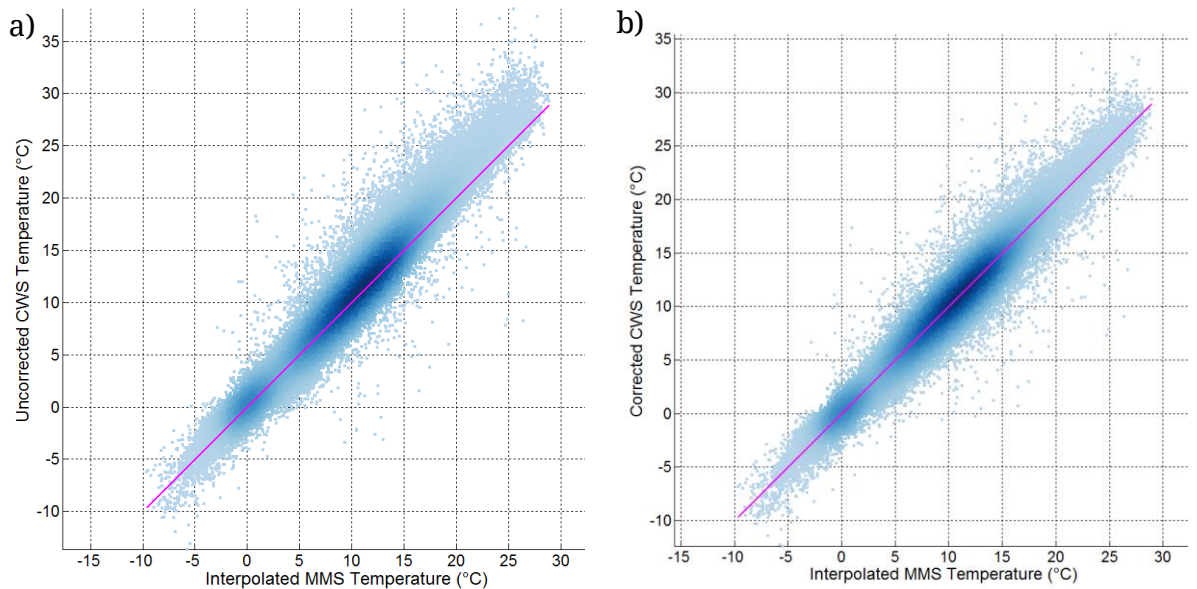


Figure 5.34. 1:1 plots of the interpolated MMS temperature observation against the a) uncorrected, and b) corrected CWS observations. Points shown are for all four 2 week case study periods. The magenta line indicates the 1:1 line.

Table 5. Mean and variance statistics for the discrepancy between CWS temperature observations and IMMS both before and after bias correction.

	Mean Discrepancy (°C)		Standard Deviation of Discrepancy (°C)	
	Before	After	Before	After
All periods	+0.54	-0.03	1.32	1.06
Autumn	+0.26	-0.07	1.18	1.04
Winter	+0.15	-0.02	0.84	0.72
Spring	+0.71	-0.02	1.38	1.15
Summer	+0.90	-0.02	1.53	1.23

Removing the systematic bias, and thus producing a virtually zero mean discrepancy, explains much of the decrease in the standard deviation of this discrepancy. However, it does not explain the full decrease. The additional decrease is a result of having applied a bias correction specific to each station and each timestep.

It is important to investigate the impact the correction has per station, as shown in Figure 5.35, as well as through time, Figure 5.36. Note that in comparison to corresponding figures that showed the uncorrected data (Figure 5.1 & Figure 5.2) the vast majority of stations that displayed an overall warm bias now exhibit a median temperature difference much closer to zero. Despite the correction there are still significant differences, for example many of the red markers in Figure 5.35 exceed ± 5

°C. This was not the case when the interpolation model was verified using withheld MMS observations (Figure 4.24). Natural spatial variations are unlikely to explain such large differences alone. This indicates the initial quality control step (Section 5.6.2) is not strict enough with gross errors still evident in the data. For example, on further inspection, the station whose whiskers are over 20 °C apart appeared to be out of phase by 12 hours indicating an incorrect location or timezone. By increasing the threshold on the *Correlation* check this station's observations should be filtered out more successfully.

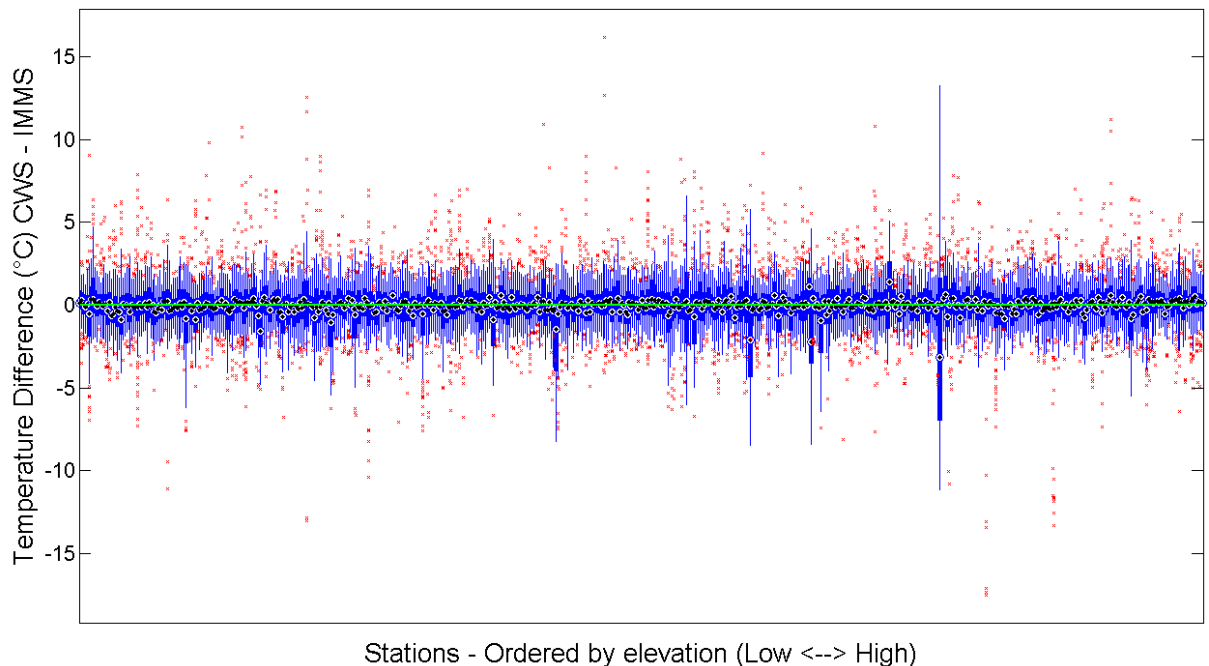


Figure 5.35. Box plot of the temperature discrepancy (CWS-IMMS) statistics for each individual station after the correction has been applied for the summer period. The black dots mark the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers ($1.5 \times$ interquartile range), and outliers are plotted individually in red. Compare with (Figure 5.1), the equivalent figure for the uncorrected CWS data.

It is interesting that Figure 5.36 still shows clear vertical banding indicating an overall diurnal pattern to the discrepancy. Whereas the uncorrected CWS data tended towards a significant warm bias during the day the corrected observations are often slightly cooler than IMMS values implying the radiation biases have been slightly overcorrected. At night there is a tendency towards a slight warm bias.

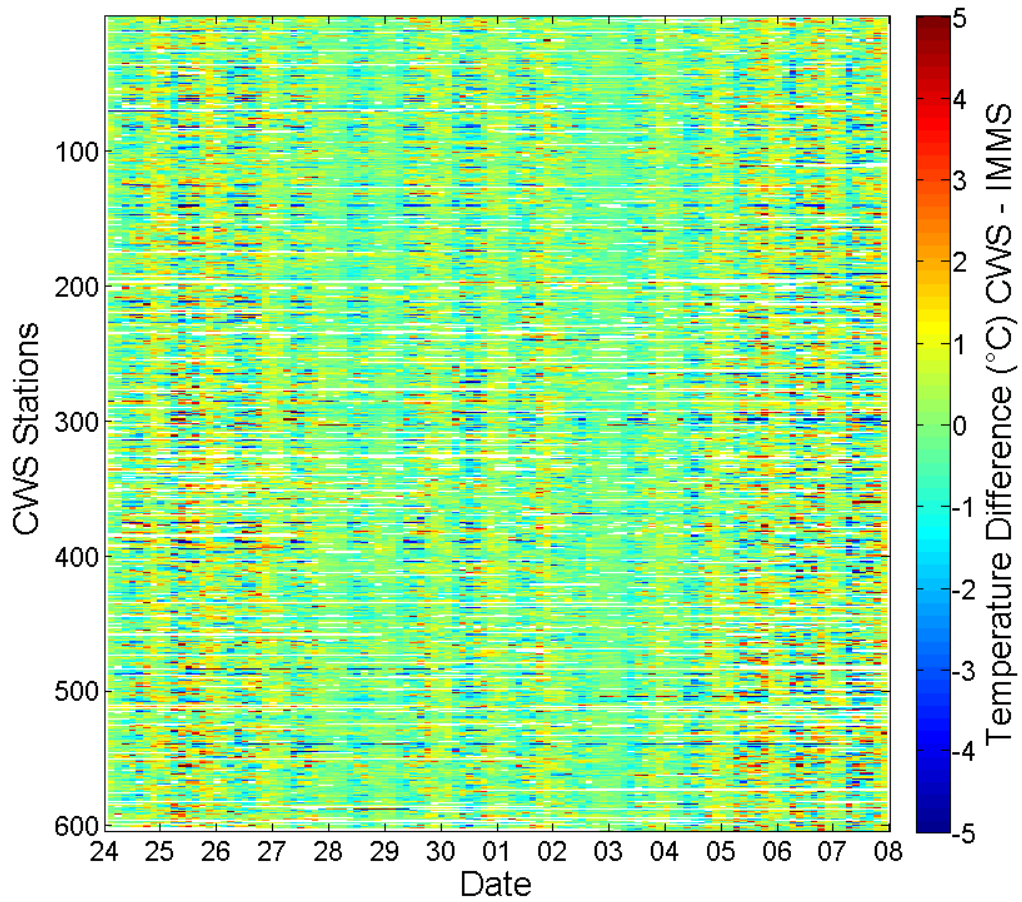


Figure 5.36. Visualisation of the difference between the corrected CWS observations and IMMS for each station (rows) and at each timestep (columns) over the summer period. Ticks on the x-axis indicate midnight at the start of that date. Compare with Figure 5.2, the equivalent figure for the uncorrected CWS observations.

It is impressive that within the short 2 week case study periods the bias correction model is able to quickly learn and correct the biases inherent to each CWS. We now look in detail at several case study stations to show the learnt calibration and radiation-induced biases through time to highlight the rate at which they are learnt.

To demonstrate clearly how the bias correction model works we begin by passing it some artificially simulated CWS data to which a known bias is added. Figure 5.37 shows the learnt calibration and radiation-induced bias for a dummy CWS station which in comparison to a simulated IMMS time series was given an artificial calibration bias of $-2\text{ }^{\circ}\text{C}$ and a radiation bias of $0.007\text{ }^{\circ}\text{C per W m}^{-2}$. The data is at 3-hourly timesteps, akin to the test set of real CWS data. From the figure it is clear that by the end of the 30 day period the learnt calibration bias has fallen close to the $-2\text{ }^{\circ}\text{C}$ level. It is important that the calibration bias is learnt gradually so that it does not react significantly to short-lived natural spatial variations, interpreting them as bias. For example, in this example it would take ~ 90 days for the learn calibration bias to drop below $-1.9\text{ }^{\circ}\text{C}$. To begin with the station's model name was set as *Unknown*

therefore equal weightings were given to every design class. As Figure 5.38 shows, by the end of the period the model has learnt that the magnitude of artificial induced radiation bias is indicative of the *Encased* design class. Note that to begin with the negative calibration bias counteracted the radiation-induced bias causing weaker daytime biases more indicative of the *Encased-Louvered* class. However once the calibration bias was learnt, and corrected for, the model was subjected to the full weight of the radiation bias and updated the design probabilities accordingly.

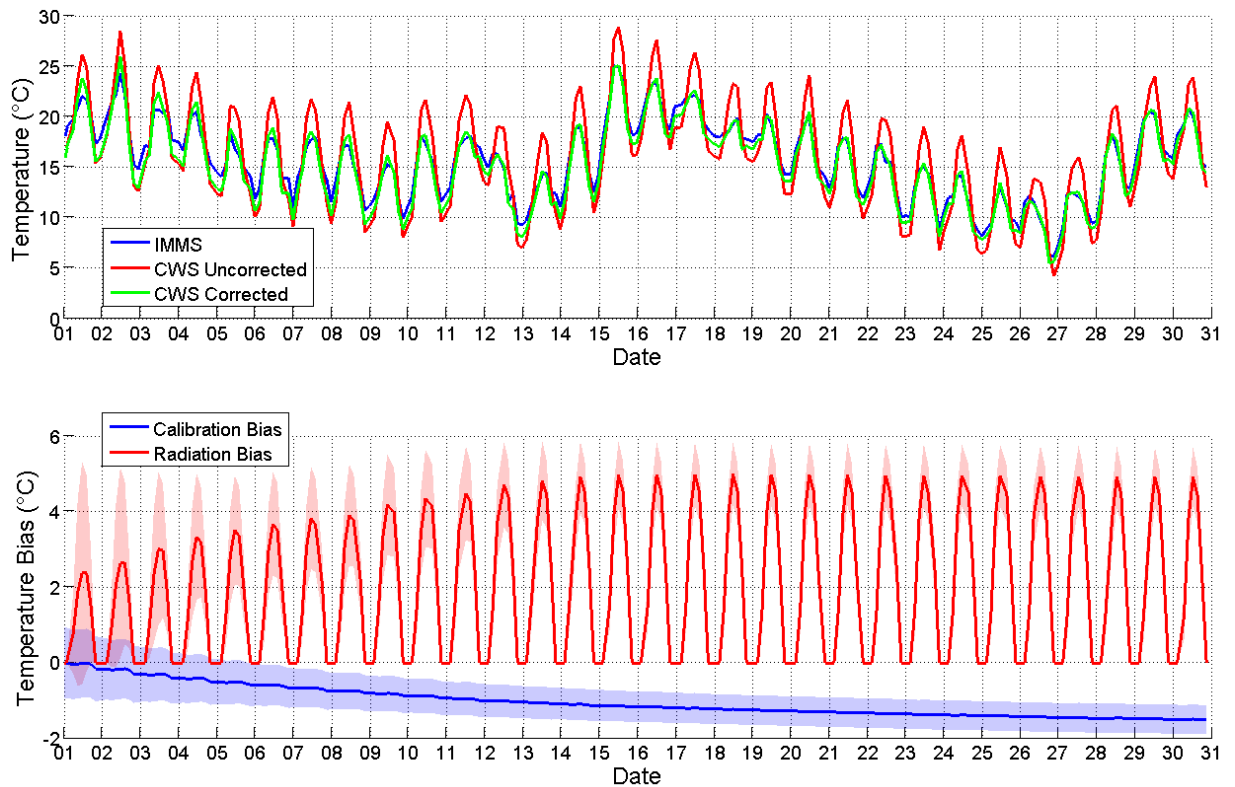


Figure 5.37. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias when the bias correction model was subjected to artificial CWS data with a calibration bias of -2°C and radiation bias of $+0.007^{\circ}\text{C per W m}^{-2}$. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.

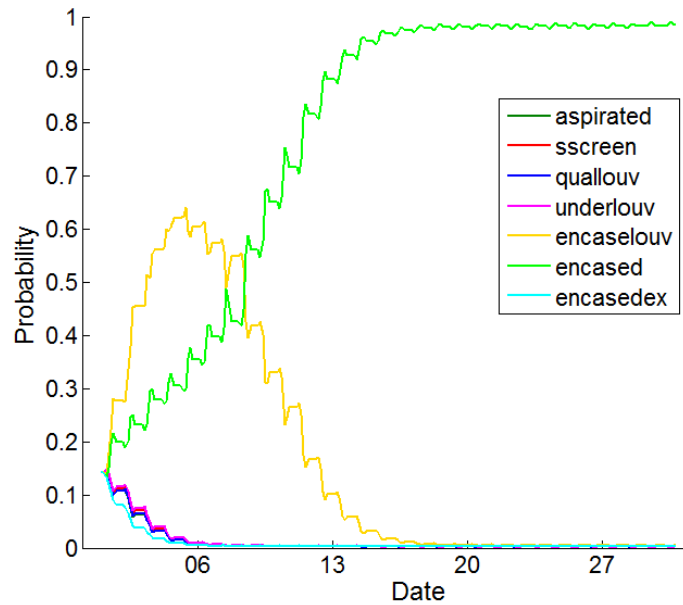


Figure 5.38. Change in design membership probabilities when the bias correction model was subjected to artificial CWS data with a calibration bias of $-2\text{ }^{\circ}\text{C}$ and radiation bias of $+0.007\text{ }^{\circ}\text{C}$ per W m^{-2} . Initially the station model was unknown therefore each class was given an equal weighting.

Having shown that the model is competent at learning artificially induced calibration and radiation biases the following figures show the same plots but for the real CWS data from WOW. Out of the hundreds of CWS stations just 3 are show below, chosen because they highlight an interesting property of the bias correction model and display features indicative of the many stations not shown. Each station's data is shown over the summer period as radiation-induced biases were most significant during this period.

Station 1

Station 1 is an example of station for which the model name was initially unknown, as in the above example. As the station continually exhibited significant daytime warm biases with respect to IMMS (Figure 5.39) the station was classified over time primarily as *Encased Louvered*. The station also exhibited a warm calibration bias which, once learnt, increased the probability that the station belonged to the *Encased-Louvered* as opposed to the *Encased* class.

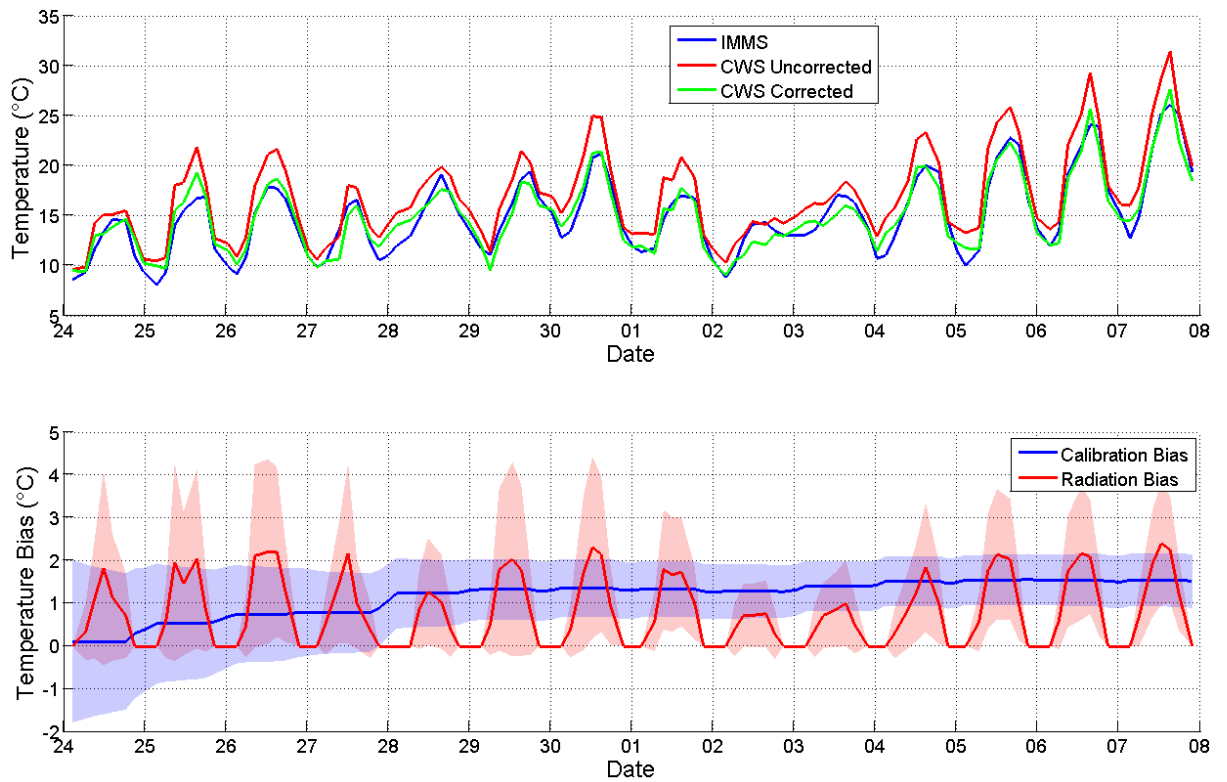


Figure 5.39. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 1. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.

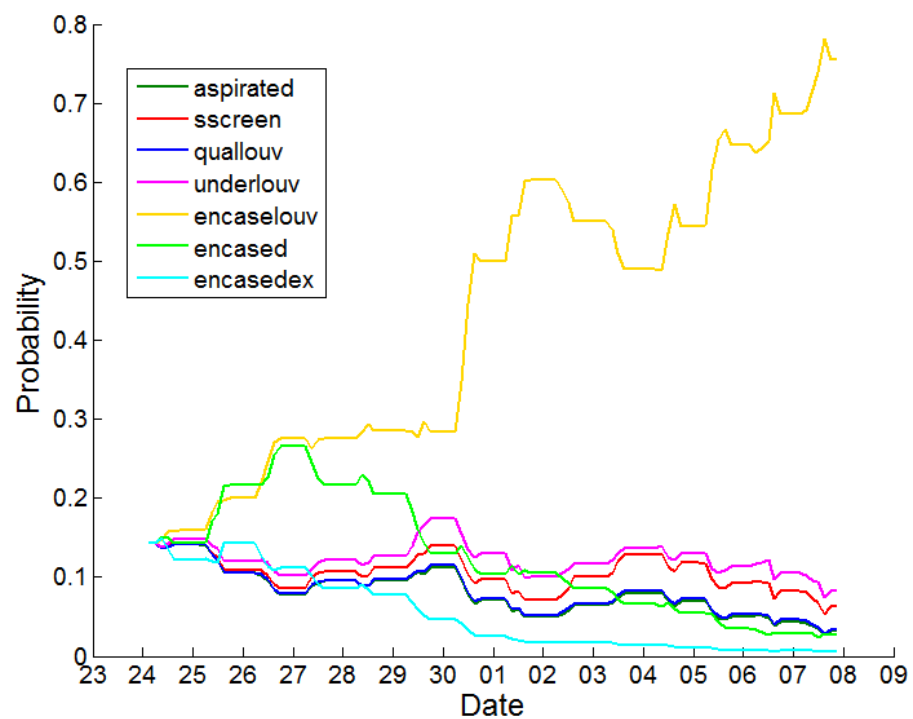


Figure 5.40. Change in design membership probabilities for case-study Station 1. Initially the station model was unknown therefore each class was given an equal weighting.

Station 2

Station 2 shows an example of station for which the user listed their model name (Fine Offset WH1080) in their metadata. As a result the station could be pre-assigned to a particular design class, in this case *Encased Louvered*. For this particular station the assigned design class was appropriate and *Encased Louvered* remained the dominant class (Figure 5.42). Relative to IMMS the station also exhibited a slight warm calibration bias a night, which was learnt by the model.

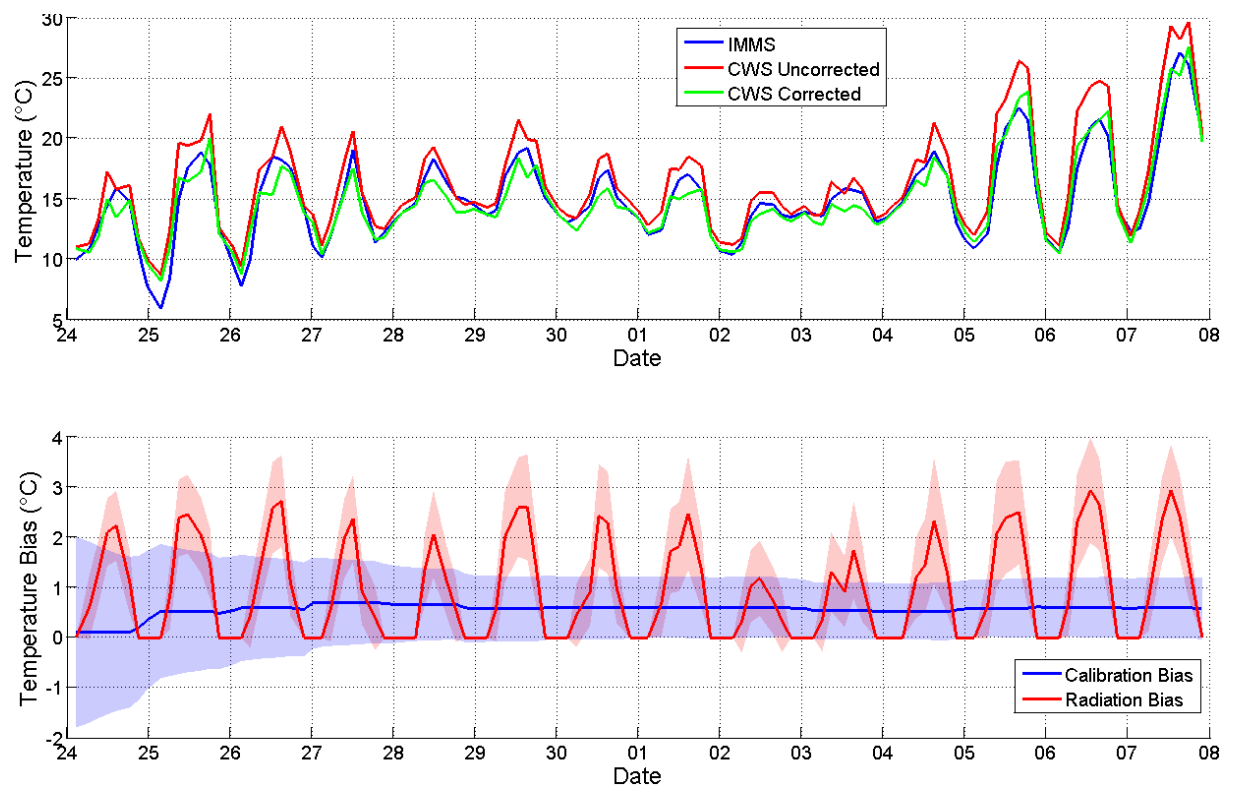


Figure 5.41. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 2. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.

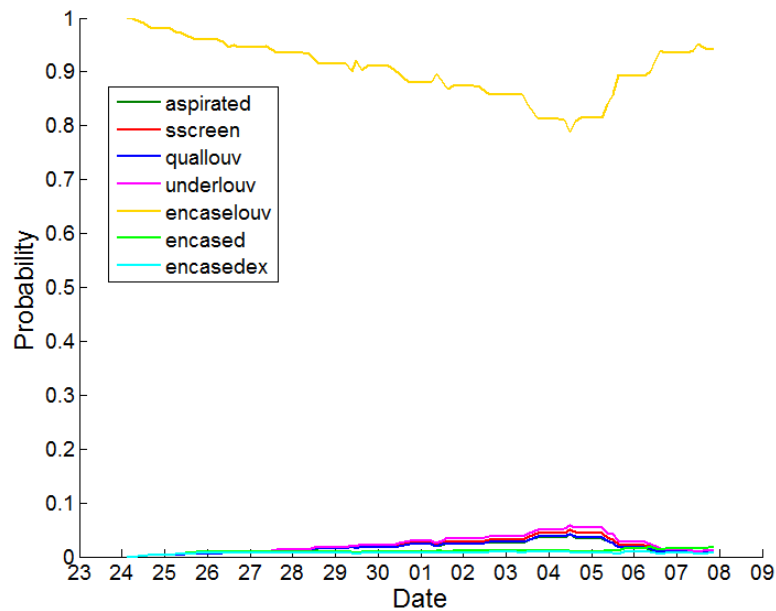


Figure 5.42. Change in design membership probabilities for case-study Station 2. Initially the design class was set as *Encased Louvered* based upon the station’s metadata.

Station 3

Station 3’s metadata implies the station model is a Davis Vantage Pro2, which performed well in the intercomparison field study with few significant biases. *A priori* the station is allocated to the *Quality Louvered* class. With respect to IMMS the station displayed minimal calibration or radiation-induced biases over the 2 week period, as expected from a *Quality Louvered* station, and therefore it remained in the same class by the end of the period. However it is interesting that the probability of the *Aspirated*, *Stevenson screen* and *Underslung Louvered* classes gradually rises over the period. Each of these classes have a similar magnitude of radiation-induced bias and later in Figure 5.50 we see that stations have a tendency to prefer the *Stevenson screen* class over *Quality Louvered*.

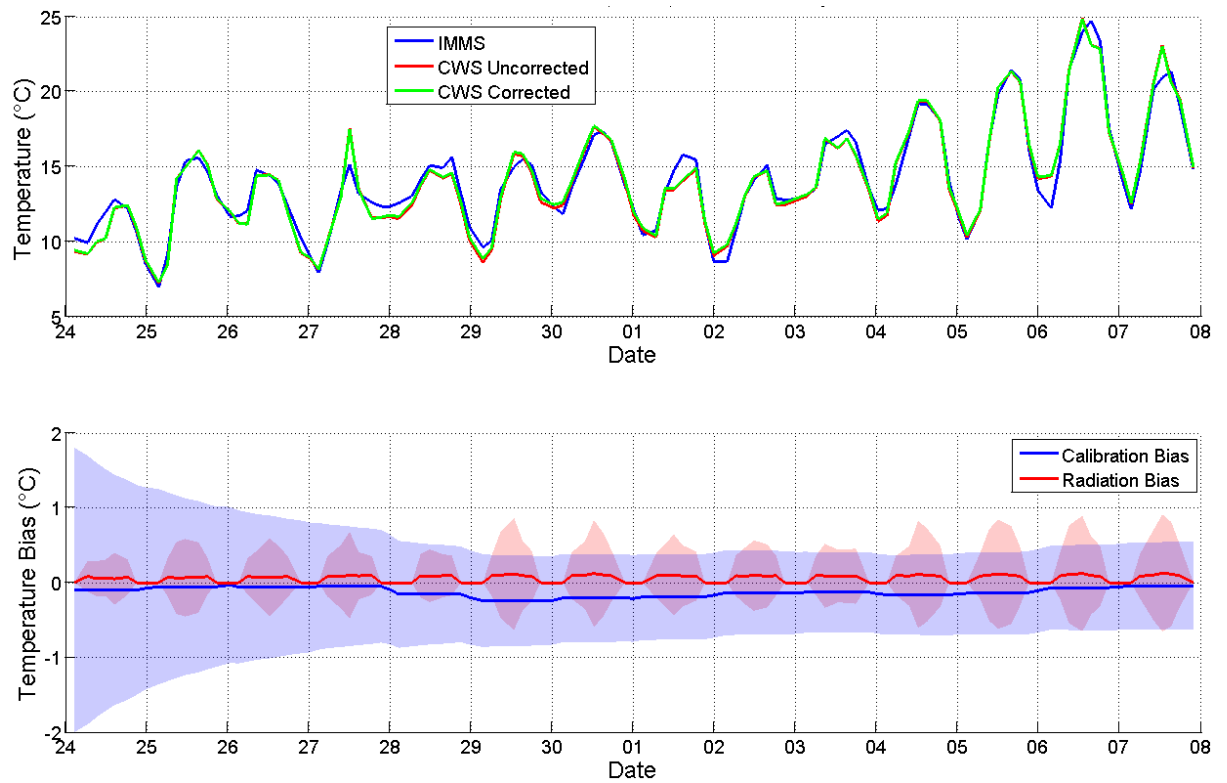


Figure 5.43. Time series of uncorrected and correct CWS data as well as the estimated calibration and radiation-induced bias for case-study Station 3. Red and blue shaded areas represent the uncertainty (± 1 s.d.) of the bias estimates.

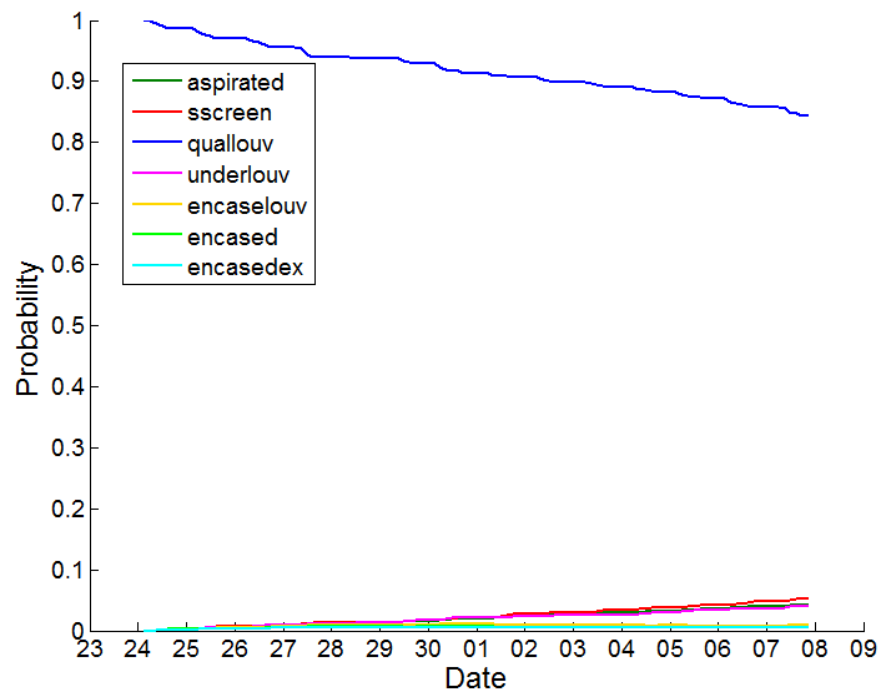


Figure 5.44. Change in design membership probabilities for case-study Station 3. Initially the design class was set as *Quality Louvered* based upon the station's metadata.

These example stations provide confidence that the bias correction model is performing as expected, and that it is able to gradually update the calibration bias and design probabilities based on the new information it receives at each timestep.

The previous time series plots showed that the learnt uncertainty associated the calibration bias estimate, $v_{s,t}^{cal}$, and radiation-induced bias estimate, $v_{s,t}^{Rad}$, evolving through time. When these two uncertainties are combined with the sensor noise, $v_{s,t}^{CWSu}$, as in Equation (32) the total uncertainty of the corrected CWS data, $v_{s,t}^{CWSc}$, is estimated. Figure 5.45 and Figure 5.46 show how this value evolved through time for all the stations during summer and winter respectively. The square root of this variance term, i.e. the standard deviation, is shown to ensure the units ($^{\circ}\text{C}$) are meaningful. It is clear to see that during the day when the model is uncertain about the degree of radiation-induced biases the overall uncertainty is therefore larger. During winter, when these radiation biases are less extreme, the uncertainty is lower. Overnight variations in the uncertainty result from changes in the calibration bias uncertainty, $v_{s,t}^{cal}$. Initially $v_{s,t=0}^{cal}$ was set as 4, i.e. a standard deviation of 2 $^{\circ}\text{C}$. Note how the model quickly reduces this overnight uncertainty.

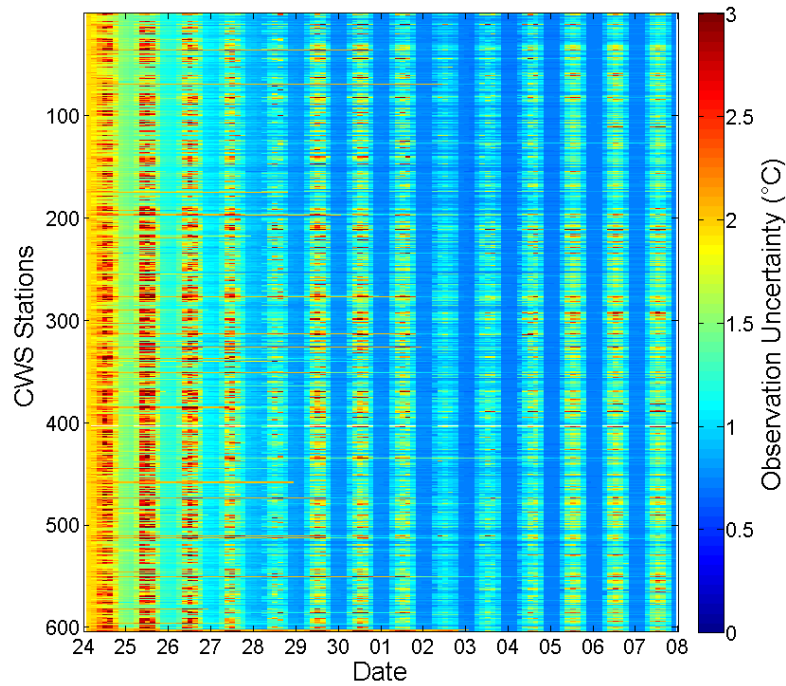


Figure 5.45. Learnt observational uncertainty for each CWS station (rows) at each time (columns) during the summer period. Shown as the standard deviation, not the variance, i.e.

$$\sqrt{v_{s,t}^{CWSc}}.$$

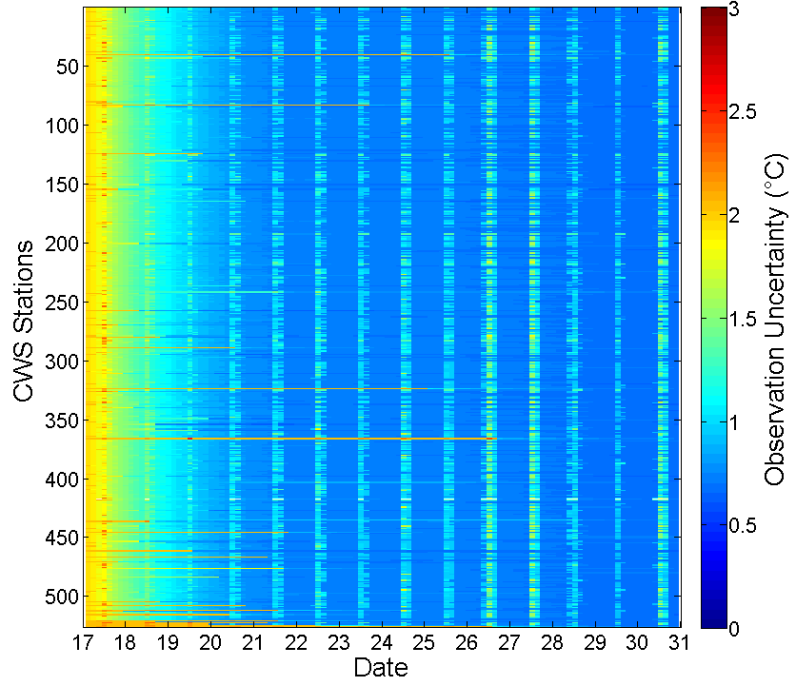


Figure 5.46. Learnt observational uncertainty for each CWS station (rows) at each time (columns) during the winter period. Shown as the standard deviation, i.e. $\sqrt{v_{s,t}^{CWSc}}$.

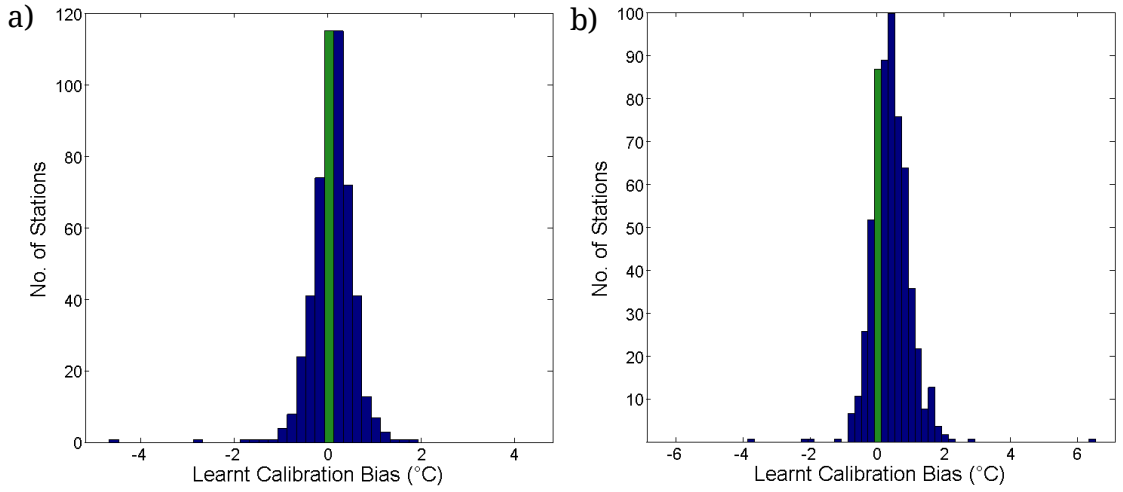


Figure 5.47. Distribution of the learnt calibration bias mean terms at the final timestep of the a) winter and b) summer periods.

Figure 5.47 shows the mean calibration bias estimate, $\mu_{s,t}^{Cal}$, for not just a single station, as the previous plots have shown, but for every station. We show its final value, so that the model has used a maximum of 2 weeks' worth of data to learn it. It is interesting that during the summer the majority of stations are allocated a positive calibration bias whereas in winter the negative/positive split is more equal. Given that $\mu_{s,t}^{Cal}$ is only updated at night it should not pick up any radiation-induced warm biases experienced by many stations during the day which could otherwise explain such an

effect. Possible explanations are that the air inside the CWS thermistor housings may remain warmer than the surrounding ambient air well into the night after a day of strong insolation. If this happened regularly this would be interpreted as a positive calibration bias, but such a phenomenon was not clearly evident within the field study data (Section 3.2). A second explanation is that because CWS, relative to MMS, are frequently located in sheltered urban locations (Figure 5.25) their siting may be responsible for the positive bias. However, in Section 5.5.2 there was little evidence that siting had a significant impact.

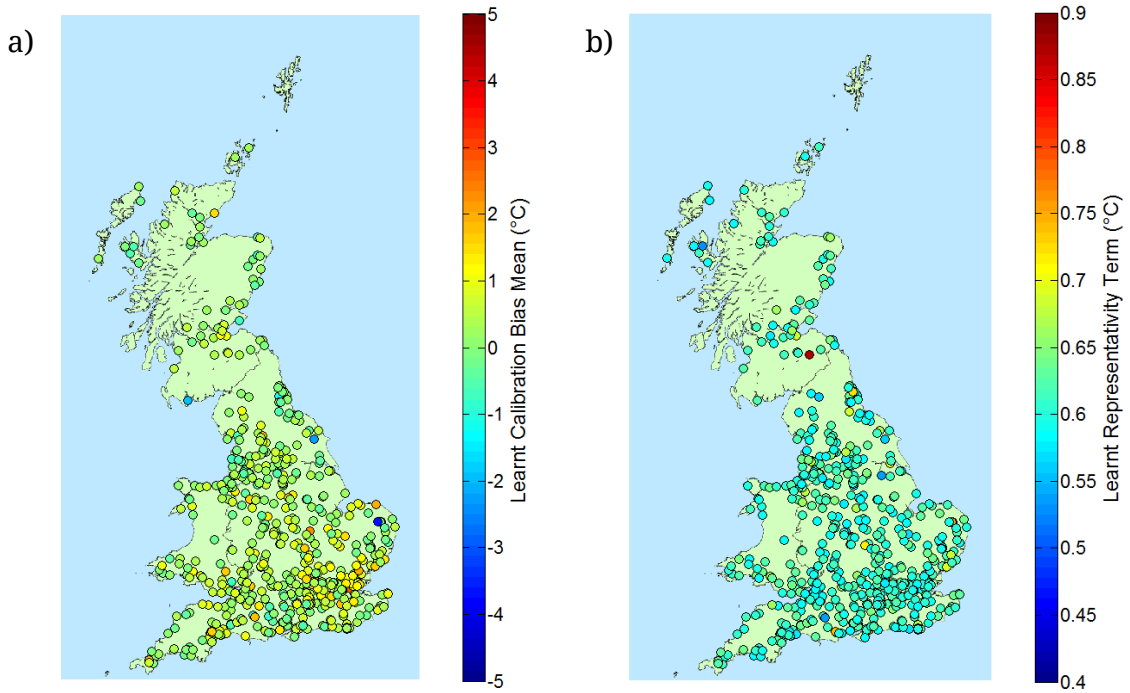


Figure 5.48. Learnt calibration mean, $\mu_{s,t}^{cal}$, and representativity term, shown as the standard deviation, i.e. $\sqrt{v_{s,t}^{CWSc}}$, at the end of the summer period plotted spatially.

Figure 5.48 shows the learnt calibration bias mean term, $\mu_{s,t}^{cal}$, and variance (representativity) term, $v_{s,t}^{cal}$, plotted spatially using their values at the final timesteps. Earlier plots, e.g. Figure 5.41, showed that these terms were relatively stable by the end of the 2 week period and thus the values shown in the spatial plots above are a fair representation of their value over most of the period. It is reassuring to see that there is no obvious spatial correlation to the mean calibration biases. For the representativity term however we would expect signs of spatial coherency as the term should be influenced by synoptic weather conditions and land cover types which vary spatially across the country. However the figures show virtually no correlation, in fact $v_{s,t}^{cal}$ barely varies across the country.

Previous plots in this section have shown the how the design probabilities change for individual stations. Figure 5.49 summaries how they change over the full CWS network. It counts which design types had the highest probabilities both at the start of the period when they were informed purely from the metadata compared with the most probable design type assigned by the bias correction model by the end of the period. For a large proportion of stations the design type at the start matches that at the end, either because the initial allocation was correct or because there has not been enough evidence within the short 2 week period for it to change. It is encouraging to see that for those stations that began as *Unknown*, a significant number were allocated to each of the bottom 4 design classes shown in Figure 5.49. It is unsurprising to see the *Encased-Louvered* class allocated to the greatest number of stations as this class contains the most common model of station, the Fine Offset WH1080 (see Figure 2.6 for a summary of station ownership).

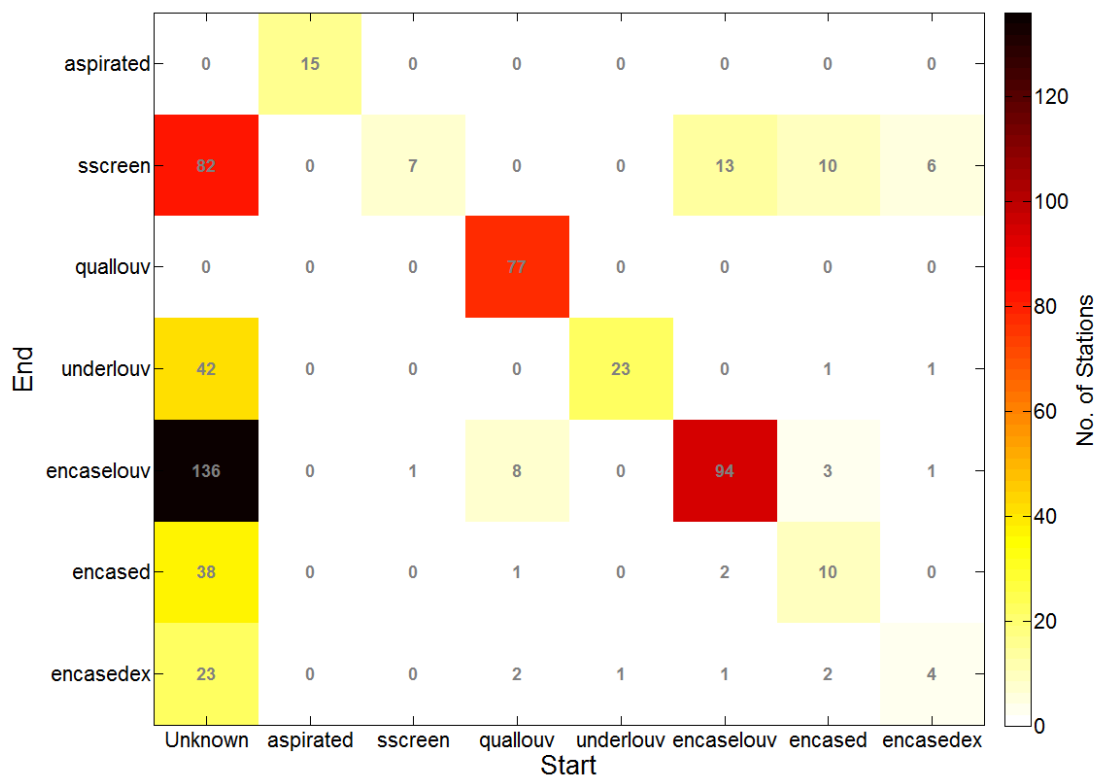


Figure 5.49. Number of stations allocated to each design class at the start (columns) vs the number at the end (rows) of the summer period. At the start the design class allocation were based upon the user’s metadata.

The results in Figure 5.50 differ from those in Figure 5.49 because before the model was run this second time, all prior knowledge of the station type derived from the metadata was ignored. Instead we allocate an equal probability to every class. The metadata-assigned classes are still shown to test whether our bias correction model, with no prior knowledge, assigns stations to the same design class as the metadata

would imply. 54 stations whose metadata implied they belonged to the *Encased Louvered* class were indeed allocated to this class by the model. However for most other design classes the agreement was less strong. In Figure 5.50 and in Figure 5.49 it is clear that for stations that display minimal radiation-induced biases the *Stevenson screen* class is favoured over the *Quality Louvered* and *Aspirated* classes. From Figure 5.18 it is clear that these 3 classes all exhibit a very similar relationship between radiation and temperature bias, i.e. there is virtually no dependency. As was evident in Figure 5.44 – when there is no strong evidence that a station belongs to a given class the model gradually shares the probabilities between the multiple classes that best match the evidence; in this case *Stevenson screen*, *Quality Louvered* and *Aspirated*. As the *Stevenson screen* is seen as the professional standard its covariance matrix for the regression coefficients used zero values on the off-diagonals. It is therefore favoured slightly more than the other two classes with non-zero off diagonals because it has slightly higher uncertainty, and thus appears as the dominant class in the figures here. As these 3 classes are so similar, our experience suggests we may be better off combining them into a single class.

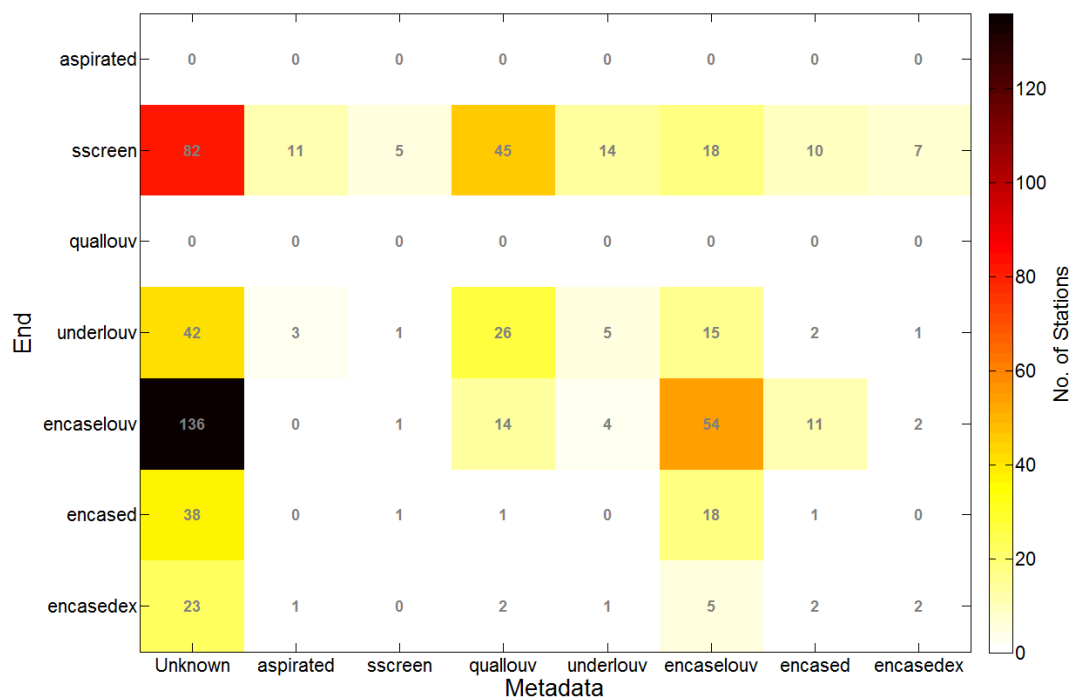


Figure 5.50. Number of stations allocated to each design class at the start (columns) vs the number at the end (rows) over the summer period. Unlike Figure 5.49 each station was allocated an equal probability to each design class at the start. The metadata classes are still shown to assess whether the learnt classes at the end match the metadata.

5.7.2. Interpolating with corrected CWS data

In this section, the corrected CWS data is fed back into the temperature interpolation model to assess whether the additional data can improve its cross-validation accuracy over using MMS data alone. We also show that if the CWS data is used without any correction it can have a detrimental impact on the model.

Although the same Bayesian linear regression model, as detailed in Section 4.3, is used for this second run of the interpolation model there are some key changes to acknowledge. Whereas previously the interpolation model was used to estimate the temperature at CWS locations, it now predicts at test MMS stations. These test MMS stations are removed from the outset so that they have no influence on IMMS values and therefore have no impact on the learnt bias corrections for the CWS stations. 10-fold cross-validation is used to iterate through different sets of test MMS stations. The second key difference is that the forecast step is no longer used. Instead of propagating the posterior for the regression coefficients forward from the last timestep to act as the prior, as in the first run, a loose prior is instead used so that the regression coefficients are essentially entirely learnt afresh at each timestep using both the CWS and MMS data. Given that the bias correction model outputs uncertainty estimates for each corrected CWS observation, $v_{s,t}^{CWS_c}$, we weight their impact in the interpolation model by this uncertainty, so that a station with a large uncertainty has little influence on the learnt regression coefficients. To do this the update step, Equations (6) and (7), has been altered so that the uncertainty inversely weights each station's impact on the learnt regression coefficient mean, $\mu_{\beta,t}$, and covariance matrix values, $\Sigma_{\beta,t}$ (Equations (33),(34) & (35)).

$$\Sigma_{\beta,t}^{-1} = X^T \omega X + \tilde{\Sigma}_{\beta,t}^{-1} + \Gamma \quad (33)$$

$$\mu_{\beta,t} = \Sigma_{\beta,t} \left(\tilde{\Sigma}_{\beta,t}^{-1} \tilde{\mu}_{\beta,t} + X^T \omega t \right) \quad (34)$$

where ω specifies the weighting given to each station, specified as a diagonal matrix. These weightings are inversely proportional to the observational uncertainty, $v_{s,t}$, assigned to each station at each timestep.

$$\omega = \frac{1}{v_{s,t}} \quad (35)$$

At present, the bias correction model only produces observational uncertainty estimates for CWS stations; i.e. for these stations $v_{s,t}$ is equivalent to $v_{s,t}^{CWS}$ from Equation (32). However, as both CWS and MMS data is used in this second run we must also assign an uncertainty value, $v_{s,t}^{MMS}$, to the MMS observations. From the RMSE plots, Figure 4.21, the average RMSE is in the order of 0.8 implying a total residual variance around $0.6 \text{ }^\circ\text{C}^2$. Given that we ascribe $0.2 \text{ }^\circ\text{C}^2$ to instrumental noise then $\sim 0.4 \text{ }^\circ\text{C}^2$ must come from representativity errors. As we attempt to quantify representativity errors for the CWS stations it is important that the MMS stations also include such an estimate. Therefore all MMS stations were ascribed an uncertainty of $0.6 \text{ }^\circ\text{C}^2$. This is relatively ad hoc – in the future it would be desirable to dynamically adjust and learn this uncertainty term as we do for the CWS stations; i.e. by learning $v_{s,t}^{Cal}$. We also run the interpolation model using the uncorrected CWS data for comparison (Figure 5.51). For these stations $v_{s,t}$ is also set as $0.6 \text{ }^\circ\text{C}^2$.

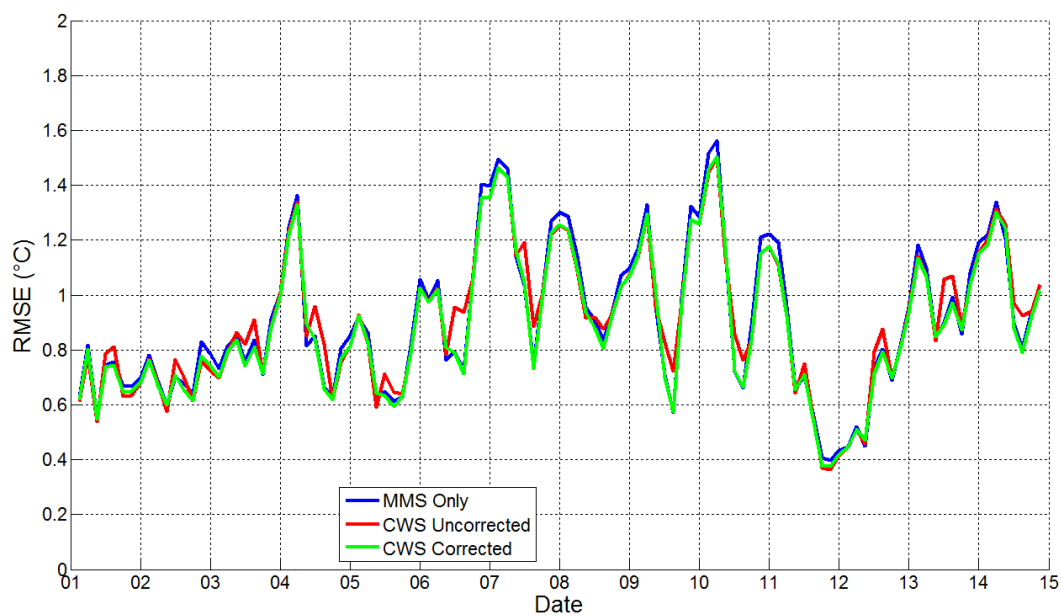
Another difference is that unlike the first run the model uncertainty, σ_ϵ^2 , does not propagate temporally. Instead a fixed value is used at every timestep. This is also assigned an ad hoc value of $\sigma_\epsilon^2 = 0.6 \text{ }^\circ\text{C}^2$, believed to represent the sum of the typical observation and representativity variance for MMS stations. Equation (11) is therefore replaced by Equation (36). Using this value ensures the model performs well probabilistically, setting this value any larger causes the model to become over-confident about its predictions; conversely a smaller value leads to an under-confident model:

$$\Sigma_{y,t} = \sigma_\epsilon^2 (1 + (X \Sigma_{\beta,t} X^\top)). \quad (36)$$

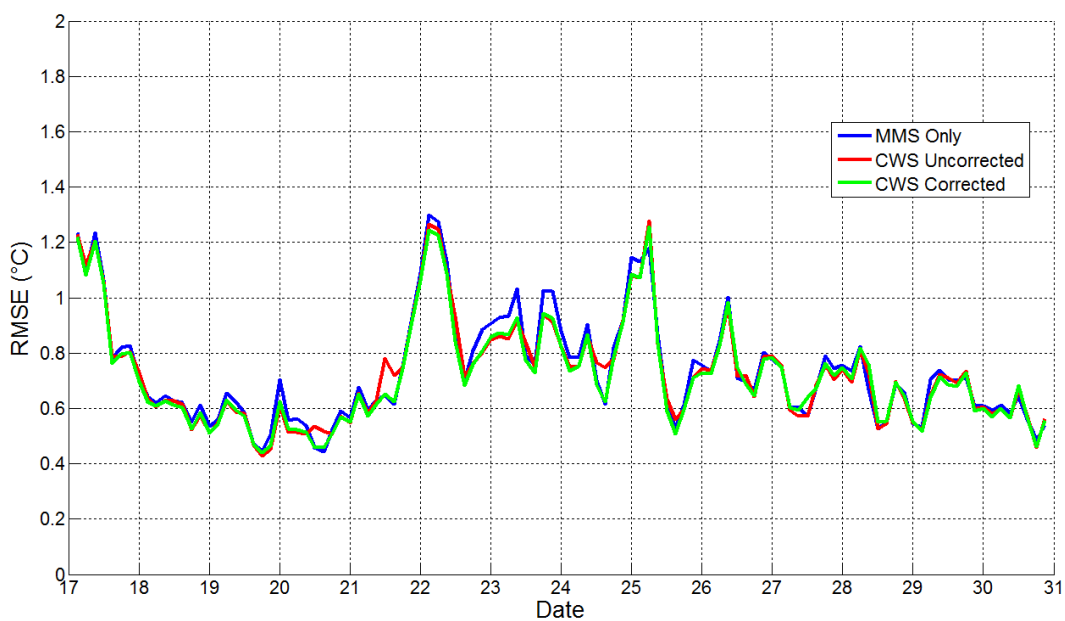
To assess fairly the impact that incorporating corrected CWS data has on the interpolation model error, this second version of the interpolation model is run 3 times, firstly with MMS data only, secondly with MMS data and uncorrected CWS data, and finally with both MMS data and corrected CWS data. From Figure 5.51 it is clear that when uncorrected data enters the model the accuracy falls most significantly in spring and summer when radiation-induced bias are at their greatest. There is a clear benefit of using corrected CWS data vs uncorrected data; however the corrected CWS data provides little if any benefit over using MMS data only. This implies that for the scale at which the interpolation model is representative, the MMS data already does a sufficient job with little room for improvement, meaning that the CWS adds no benefit. It is also worth noting that MMS data might not validate the CWS due to a lack of ‘urban’ MMS stations. It is in urban locations where the addition of CWS may provide the greatest benefit. There may be other applications however that can

extract greater value from the CWS data. For example, the corrected data may prove valuable within a data assimilation scheme capable of resolving higher spatial resolutions.

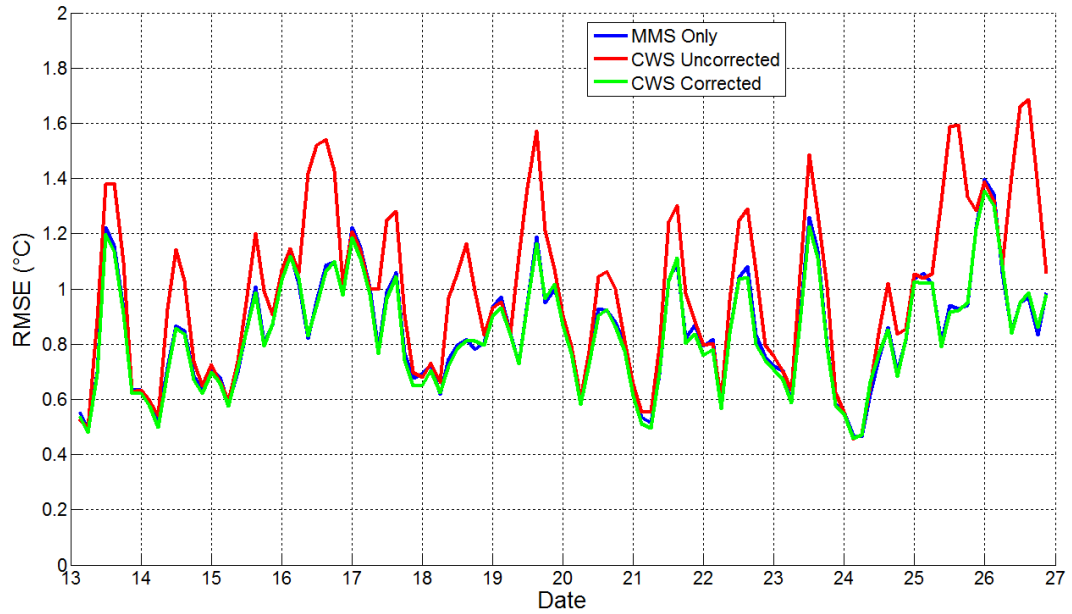
a) Autumn



b) Winter



c) Spring



d) Summer

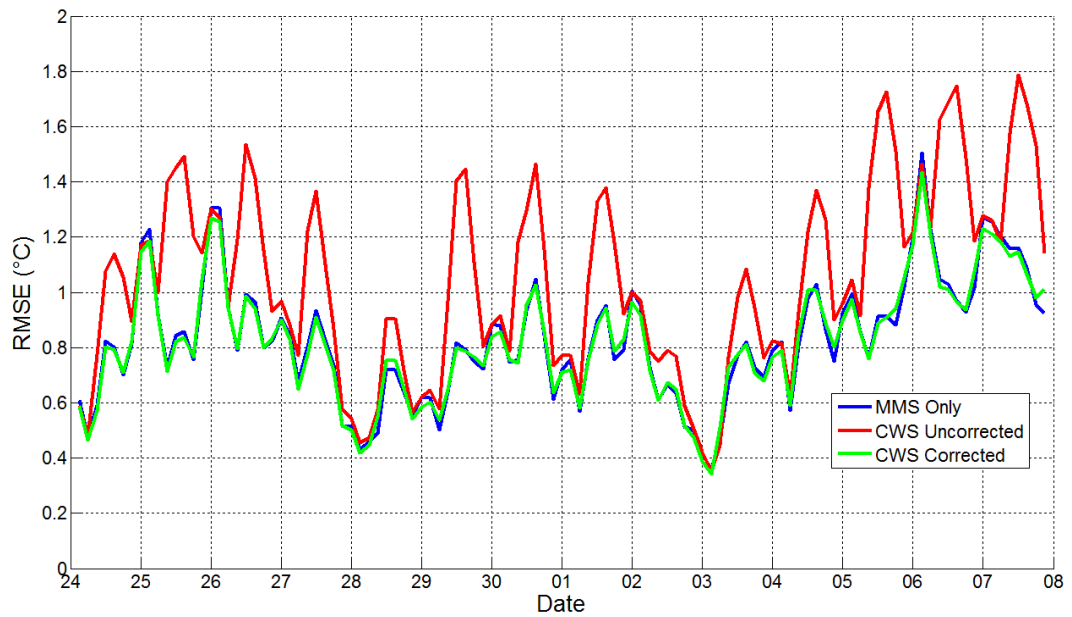


Figure 5.51. Time series of cross-validation RMSE of the temperature interpolation model run under 3 scenarios: with MMS data only, with MMS data and uncorrected CWS data, with MMS data and corrected CWS data. Shown for each two week case study period: a) Autumn, b) Winter, c) Spring, d) Summer.

A coverage plot, Figure 5.52, is included to show that the interpolation model still validates well probabilistically; i.e. on average the uncertainty values that accompany the predictions at the test MMS stations fairly reflect the prediction errors.

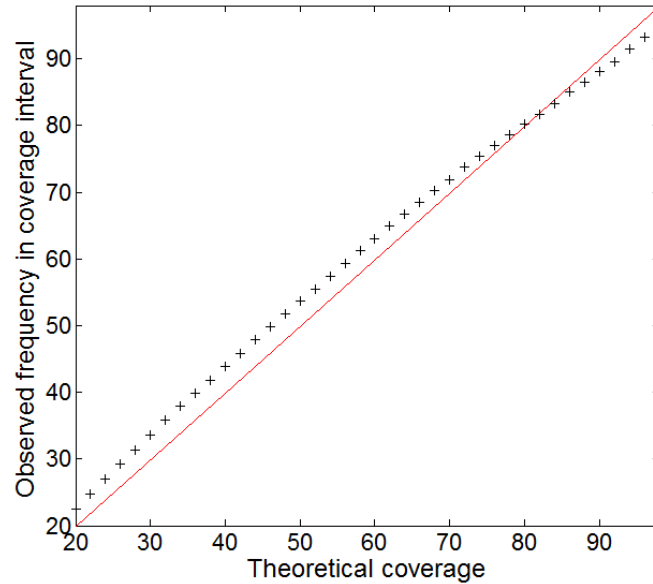


Figure 5.52. Coverage plot for when the temperature interpolation model was run with MMS data and corrected CWS data over the summer period. Verified against withheld MMS station observations using 10-fold cross-validation. It plots the theoretical centred confidence interval against the observed frequency.

5.8. Summary

Within this chapter we have conceptualised, designed and tested a model capable of learning biases within CWS temperature data; a model built upon explicitly modelling calibration and radiation-induced biases. The standout feature of this model is its Bayesian approach to learning these biases over time so that they can be accurately estimated, with a quantified uncertainty, without correcting for short-lived natural spatial variations which we wish to capture.

Crucially, we were able to test the model with real CWS data. Having investigated the impact of the model's bias correction both on the dataset as a whole and looking at individual case study stations, there are clear signs that our approach is capable of significantly reducing the biases inherent to the data. The model also appears to attach realistic uncertainty estimates to each observation; estimates that would be informative to any potential users of the data when deciding if the observations are fit for purpose.

It is interesting that the corrected CWS data added little value when included alongside MMS data in our temperature interpolation model. It appears that in this application the MMS data alone is sufficient; however, other applications of the corrected CWS data should be tested in future work to better leverage their potential value.

Although the focus was on temperature observations, several of the techniques introduced here should be applicable to other variables. The Bayesian framework could be easily applied to other variables. The radiation interpolation model developed here would also be of use if the model is used to learn bias in dew point observations, which also have a dependency on radiation. The technique of assigning stations to different design classes based up their bias characteristics should also work well for dew point and also for precipitation observations as the size and shape of rain gauges play a role in determining a station's over or under catch.

6. Conclusion

This project has successfully combined a variety of techniques, data, and fields of research to demonstrate an approach for quantifying bias and uncertainty within citizen weather data. Below we detail the project's key successes and how they contribute to the field of citizen meteorology. We note areas that still require further work, such as how such a system could be implemented operationally. We also make a number of suggestions based upon what has been learnt during this project; suggestions that should help improve the quality of CWS data in the future.

6.1. Contributions

The first major achievement of the project was to successfully identify and parameterise bias within CWS data, with a particular focus on being able to parameterise radiation-induced biases with interpolated professional global radiation observations. We have shown that for stations with inadequate radiation shielding, using such a parameterisation can dramatically reduce the mean bias and residual variance of their observations. The decision to conduct a year-long intercomparison study of popular CWS proved invaluable in this process, helping to quantify the magnitude of the CWS bias we would expect in real CWS data. Such an approach could be easily implemented by others wishing to correct CWS temperature observations they suspect have significant radiation-induced biases, assuming reliable estimates of incoming radiation are available.

The bias correction model relies on temperature and radiation estimates at the CWS locations. Here we successfully demonstrated that a Bayesian linear regression model can be used to interpolate professional MMS temperature and radiation observations to provide such estimates. For studies where quantifying the uncertainty of these interpolated values is important this approach provides a robust solution. The benefit for this study was that the uncertainties could be passed to our bias correction model and propagated through to the final uncertainty of the corrected CWS observations. Much of the success of the interpolation models came from their ability to incorporate crucial datasets. For example, the temperature interpolation model incorporated high resolution forecast model output, whereas the radiation interpolation model leveraged publicly-available satellite imagery.

A primary success of this project has been to develop and demonstrate a complete approach for quantifying bias and associated uncertainties within CWS temperature observations. In contrast with previous quality control techniques, we were able to

demonstrate an approach in which data was not simply flagged as erroneous and discarded, but instead explicitly model the calibration and radiation-induced biases such that even biases over several degrees Celsius could be corrected for, accounting for the important uncertainties, leaving the data available for use. This lays the foundation for an operational implementation of such a system, which should ultimately discard fewer CWS observations and allow end users to assess the appropriateness of the data themselves using the assigned uncertainty estimates. We have also demonstrated the dangers of using uncorrected CWS data, with uncorrected WOW data having a detrimental impact on our temperature interpolation model.

Many CWS stations submit at sub-hourly intervals and in parts of the country the spatial density of stations far exceeds that of professional networks. With a reliable quality control procedure in place, such as the one developed herein, this abundance of data stands a real chance of improving the performance of suitable applications. For example, having demonstrated an approach for quantifying the observation uncertainty, CWS data could be fed into existing high resolution data assimilation schemes with the aim of improving the initial conditions in numerical weather forecast models. With more realistic initial conditions comes the possibility of more accurate forecasts.

6.2. Implementing operationally

The aim of this project was to demonstrate an approach for quantifying bias and uncertainty in CWS data, not to implement an operational system. The UK Met Office worked closely with this project with the vision that once developed, such a system could be made operational, processing data submitted to WOW. Having presented our work at the Met Office in February 2015 the feedback was positive. However several, non-trivial, steps remain to be addressed before such a system could be run operationally, particularly in near real-time. For example, at each timestep it would be necessary to request up-to-date professional and citizen data along with satellite images and model output. These are essentially IT issues and as the ‘Internet of Things’ continues to grow and APIs for handing data requests become common place this should become more less of an issue. We also showed that once the required data has been retrieved the processing time of all the models is entirely suitable for real-time use. In this project *MATLAB* was used to process and analyse the data; operationally a language such as *Python* may be more suitable. In this project we ran the complete model at 3-hourly timesteps. Doing so fails to leverage the high temporal resolution of CWS data, with most stations submitting at sub-hourly intervals. However, as satellite data and model output is commonly only available at hourly

timesteps a challenge arises if we wish to interpolate professional temperature and radiation observations at sub-hourly intervals. It is likely that such a model would need to incorporate temporal interpolation as well as spatial interpolation. A potential application of CWS data within the Met Office is to feed it into pre-existing data assimilation schemes, in particular its high resolution configurations (Dixon, et al., 2009). In order for our CWS uncertainty estimates to propagate into such a data assimilation scheme significant changes would need to be made to the code to enable the observation error covariance matrix to update in near real time. Alternatively our approach may be more suited to post-processing (Moseley, 2011). In many respects the post-processing team is better suited to implementing such an approach, as they commonly process both model output data and observations with the aim of correcting for biases.

6.3. Advice

From the results of this project we are able to make a number of suggestions; suggestions that can help improve the quality of CWS data in the future and ensure that the data is used sensibly. Below we list suggestions for the citizen observers themselves, for station manufactures, for data hubs such as WOW, and for those wishing to incorporate CWS data into their application.

For citizens – In this study we have highlighted the roles that station type and station siting play on inducing errors in CWS observations. When metadata is available detailing these attributes, it becomes much easier to get a handle on the biases a CWS is likely to display. It is therefore vital that citizen observers complete as much relevant metadata as they can. Citizens can also take steps to reduce representativity errors by improving the exposure of their station, although we acknowledge that in many gardens an ideal siting is difficult to find. The field study results also highlight that citizens should take care when choosing their station, in particular that the station's radiation shielding is adequately ventilated with effective shielding from incoming and outgoing radiation. Here we saw that aspirated stations, and stations with a louvered design, performed best.

For manufacturers – This study highlighted several CWS design features that can induce biases and should be avoided. Firstly the radiation shielding used to protect the thermistors from direct radiation should be white and allow sufficient ventilation. Louvered and aspirated designs work well. There is an argument with the latter that the aspiration should function both day and night. This is as there were signs that by altering the shielding design to accommodate aspiration it also caused slight warm biases when the aspiration was switched off. Stations in which the thermistor was

mounted on a circuit board encased within a plastic enclosure tended to exhibit dramatic radiation-induced biases; therefore such designs should be avoided. It should also be easier for citizens to calibrate their thermistors. A suggestion would be to sit the thermistors at the end of a length of wire rather than soldering them directly to the circuit board so that a water bath could be used for calibration. We also saw that for integrated sensor suits, such as Davis' Vantage Vue, it is virtually impossible to find a siting where the height and exposure is ideal for every sensor. A modular design in which the different components can be mounted independently would overcome this. The rain gauges should also be larger, circular, with deeper sides and larger tipping buckets within, as many stations exhibited a significant undercatch in comparison to standard gauges. Hopefully through publications such as Jenkins (2014) and Burt (2013) citizens will become more aware of the errors inherent to certain brands of CWS and will demand designs with improved accuracy from manufacturers.

For data hubs – There are a number of improvements that could be made to the websites that accept, archive, and display CWS data. At present each data hub lists different metadata attributes. Ideally each hub should use a standardised metadata form in which key properties such as station type and siting are compulsory when setting up a station. With a consistent set of station model names it would be much easier to apply the appropriate radiation-induced bias correction when a new station signs up. Alternatively users could choose from a list of design classes, such as the 7 listed in Section 5.4.2. It is important that these websites continue to offer advice on how to set up a station - for example, advising what key design features to look for when buying a new CWS and what the best practices are for siting a station and how to maintain it. Data hubs may also wish to consider implementing a quality control procedure such as that demonstrated in this project so that every observation they display has an accompanying bias and uncertainty estimate. This would not only help inform data users of the data's accuracy, but could also warn the station's owner of biases that they may have previously been unaware of. Also, assuming the permission of the citizen observer has been given, more observations, and even metadata, should be made available through APIs. This would improve the accessibility of the data, and negate the need for web scraping techniques as used in this study. With easier access comes the possibility of greater research into, and application of, CWS data.

For data users – This study has shown that CWS data can contain significant biases, which if used uncorrected can have a detrimental impact on the applications in which they are used. Data users should therefore consider applying an operational version of a quality control system such as the one demonstrated here. This would ensure that

gross errors are removed, biases can be corrected for, and the uncertainty estimates attached to the data can inform how the data is used and how much confidence can be placed upon it.

6.4. Further work

Despite the successes of this project there remain a variety of challenges that require further work.

In this study the performance of our quality control system was tested by adding the corrected CWS data back into the temperature interpolation model. Here we saw little benefit over using the professional data alone. It is likely that other applications would be able to better leverage the increased spatial resolution that CWS data can provide. For example it would be interesting to add the corrected data into high resolution data assimilation schemes to assess the impact on forecast skill. This study also only uses 2 weeks' worth of continuous data at a time due to time and computational limitations. It would be interesting to see how the learnt calibration bias and design class memberships evolve over much longer timescales, e.g. over a complete year.

This project only focused on correcting temperature observations. There is however a lot of potential value in CWS humidity and precipitation observations. It is likely that the approach used for temperature would also apply well to humidity observations. Although as humidity is constrained to the limits 0-100% we would suggest modelling dew-point temperature instead for which the errors are more Gaussian. Bias correcting CWS precipitation observations would require an alternative approach. Nearby professional rain gauges could be used to buddy-check CWS gauges, potentially using radar to aid the interpolation. The field study indicated that the percentage by which a CWS gauge under or over caught remains relatively constant through time. It is possible that by interpolating professional gauge data to the CWS location this percentage could be learnt over time. It is important that a relatively long timescale is used to learn this correction so that instrumental biases can be removed without also removing short-lived spatial variations in rainfall. Longer-lived spatial variations, such as rain shadow effects, can complicate this process however.

The design type classes chosen in this study were based upon the empirical results from the field study. However, as only a limited number of stations were tested it is possible that other models of CWS available on the market display different bias characteristics. For example, they may display a different relationship with radiation or have a tendency for an 'out of the box' calibration bias. Therefore to establish

whether we need to model any other design classes further field studies could be performed on stations whose design does not closely resemble those tested here.

Quantifying representativity errors still proves to be a significant challenge. In this study the representativity term displayed very little spatial correlation. In the future this property could be enforced, as we would expect to see the representativity term react to changes in the synoptic conditions which vary across the country. There is also an argument that the learning and forgetting rate terms should be set differently for the calibration bias mean and variance terms. As the variance is used as the representativity term we would expect it show a synoptic dependency perhaps with a diurnal pattern. It should therefore update faster than the mean which is used simply to represent the instrumental calibration bias; a property that we expect to change over much longer timescales, with no synoptic dependency. The field study also suggested that the sheltering and shading of a CWS can influence the magnitude of the bias it displays. This is important as it could cause a station to be allocated to the wrong design class. This is something that could be investigated further, although accurately estimating the shading effects at a given CWS location is a very difficult task. This is partly due to errors in the exact location of the station and partly because measurements of nearby obstructions are hard to come by.

There are several other tweaks that could be made to the system. For example the clustered approach introduced in Section 4.3.3 could be further developed with the hope of improving the overall accuracy and uncertainty estimates of the temperature interpolation model. Also at present the uncertainty estimates from the radiation interpolation model are not used. In the future we would like to propagate its uncertainty through the bias correction model so that during the day it influences the final uncertainty attached to each CWS observation.

Overall, given the successes of this project, a quality control system such as the one demonstrated here has a lot of potential with plenty of room for development. Hopefully over the coming years the volume of citizen weather data will continue to grow and meteorological organisations and data users that handle such data will see the value in further developing and implementing such a system operationally.

7. References

- Alvarez, O. et al., 2014. Comparison of elevation and remote sensing derived products as auxiliary data for climate surface interpolation. *International Journal of Climatology*, Volume 34, pp. 2258-2268.
- Alves, E. & Biudes, M., 2013. Method for determining the footprint area of air temperature and relative humidity. *Acta Scientiarum*, 35(2), pp. 187-194.
- Bell, S., Cornford, D. & Bastin, L., 2013. The state of automated amateur weather observations. *Weather*, 68(2), pp. 36-41.
- Bell, S., Cornford, D. & Bastin, L., 2015. How good are citizen weather stations? Addressing a biased opinion. *Weather*, 70(3), pp. 75-84.
- Best, M., 2005. Representing urban areas within operational numerical weather prediction models. *Boundary Layer Meteorology*, Volume 114, pp. 91-109.
- Bishop, C., 2007. *Pattern Recognition and Machine Learning*. 1st ed. New York: Springer Science and Business Media.
- Blanc, P., Gschwind, B., Lefèvre, M. & Wald, L., 2011. The HelioClim Project: Surface Solar Irradiance Data for Climate Applications. *Remote Sensing*, Volume 3, pp. 343-361.
- Bohnenstengel, S., Schlünzen, K. & Beyrich, F., 2011. Representativity of in situ precipitation measurements – A case study for the LITFASS area in North-Eastern Germany. *Journal of Hydrology*, 400(3-4), pp. 387-395.
- Bosch, J., Lopez, G. & Batlles, F., 2008. Daily solar irradiation estimation over a mountainous area using artificial neural networks. *Renewable Energy*, Volume 33, pp. 1622-1628.
- Bouttier, F. & Courtier, P., 1999. Data assimilation concepts and methods. *ECMWF training course notes*.
- Brandsma, T., 2004. Parallel air temperature measurements at the KNMI-terrain in De Bilt (the Netherlands) May 2003 – April 2005. *Interim Report. KNMI-publicatie 207 HISKLM 7*, pp. 1-29.
- Bröcker, J. & Smith, L., 2007. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, 22(3), pp. 651-661.

- Bromiley, P., 2013. Products and Convolutions of Gaussian Probability Density Functions. *Tina Memo No. 2003-003*.
- Browning, K. A. et al., 2007. The convective storm initiation project. *Bull. Am. Meteorol. Soc.*, Volume 88, pp. 1939-1955.
- Browning, K. A. & Hill, F. F., 1984. Structure and evolution of a mesoscale convective system near the British Isles. *Q. J. R. Meteorol. Soc.*, Volume 110, pp. 897-913.
- Burt, S., 2009. *The Davis Instruments Vantage Pro2 wireless AWS – an independent evaluation against UK-standard meteorological instruments*. [Online] Available at: measuringtheweather.com [Accessed 16 02 2015].
- Burt, S., 2012. *The Weather Observer's Handbook*. 1st ed. New York: Cambridge University Press.
- Burt, S., 2013. *Instrument review Davis Instruments Vantage Vue AWS*. [Online] Available at: measuringtheweather.com [Accessed 16 02 2015].
- Cheung, H. K., Levermore, G. J. & Watkins, R., 2010. A low cost, easily fabricated radiation shield for temperature measurements to monitor dry bulb air temperature in built up urban areas. *Building Services Engineering Research and Technology*, 31(4), pp. 371-380.
- Clark, M., Lee, D. & Legg, T., 2014. A comparison of screen temperature as measured by two Met Office observing systems. *International Journal of Climatology*, Volume 34, pp. 2269-2277.
- Clark, M. R., 2011. An observational study of the exceptional 'Ottery St Mary' thunderstorm of 30 October 2008. *Meteorol. Appl.*, Volume 18, pp. 137-154.
- Clark, M. R., 2012. Doppler radar observations of non-occluding, cyclic vortex genesis within a long-lived tornadic storm over southern England. *Q. J. R. Meteorol. Soc.*, Volume 138, pp. 439-454.
- Courault, D. & Monestiez, P., 1999. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast France. *International Journal of Climatology*, Volume 19, pp. 365-378.
- CWOP, 2005. Weather station siting, performance and data quality guide. March 2005.

- Daly, C., 2006. Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, Volume 26, pp. 707-721.
- Daly, C. et al., 2002. A knowledge-based approach to the statistical mapping of climate. *Climate Research*, Volume 22, pp. 99-113.
- Danielson, J. & Gesch, D., 2011. Global multi-resolution terrain elevation data 2010 (GMTED2010):. *U.S. Geological Survey Open-File Report 2011-1073*, p. 26.
- Degaetano, A. & Belcher, B., 2007. Spatial Interpolation of Daily Maximum and Minimum Air Temperature Based on Meteorological Model Analyses and Independent Observations. *Journal of Applied Meteorology and Climatology*, Volume 46, pp. 1981-1992.
- DeGaetano, A. T. & Wilks, D. S., 2009. Radar-guided interpolation of climatological precipitation data. *International Journal of Climatology*, Volume 29, pp. 185-196.
- Dixon, M. et al., 2009. Impact of Data Assimilation on Forecasting Convection over the United Kingdom Using a High-Resolution Version of the Met Office Unified Model. *Monthly Weather Review*, 137(5), pp. 1562-1584.
- Eden, P., 2009. Traditional weather observing in the UK: an historical overview. *Weather*, Volume 64, pp. 239-245.
- Eden, P., 2012. Weather log - October 2012. *Weather*, 67(12).
- Eden, P., 2013a. Weather Log - January 2013. *Weather*, 68(3).
- Eden, P., 2013b. Weather Log - May 2013. *Weather*, 68(7).
- Eden, P., 2013c. Weather Log - June 2013. *Weather*, 68(8).
- Eden, P., 2013d. Weather Log - July 2013. *Weather*, 68(9).
- Fritz, S. et al., 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*, 1(3), pp. 345-354.
- Gelman, A., Carlin, J. B., Stern, H. & Rubin, D., 2003. *Bayesian Data Analysis*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
- Gibbs, A. & Su, F., 2002. On choosing and bounding probability metrics. *International Statistical Review*, 70(3), pp. 419-435.
- Green, A., 2010. From Observations to Forecasts – Part 7. A new meteorological monitoring system for the United Kingdom's Met Office. *Weather*, 65(10), pp. 272-277.

- Green, M., 1970. Effects of exposure on the catch of rain gauges. *J. Hydrol.(NZ)*, Volume 9, pp. 55-71.
- Guo, J. C. Y., Urbonas, B. & Stewart, K., 2001. Rain Catch under Wind and Vegetal Cover Effects. *Journal of Hydrologic Engineering*, Volume 6, pp. 29-33.
- Gura, T., 2013. Amateur experts. *Nature*, Volume 496, pp. 259-261.
- Gutierrez-Corea, F., Manso-Callejo, M., Moreno-Regidor, M. & Velasco-Gómez, J., 2014. Spatial Estimation of Sub-Hour Global Horizontal Irradiance Based on Official Observations and Remote Sensors. *Sensors*, Volume 14, pp. 6758-6787.
- Hamill, T., 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, Volume 129, pp. 550-560.
- Harrison, R. G., 2010. Natural ventilation effects on temperatures within Stevenson screens. *Quarterly Journal of the Royal Meteorological Society*, Volume 136, pp. 253-259.
- Harrison, R. G., 2011. Lag-time effects on a naturally ventilated large thermometer screen. *Q. J. R. Meteorol. Soc.*, Volume 137, pp. 402-408.
- Hengl, T. & Heuvelink, G. R. D., 2007. About regression-kriging: From equations to case studies. *Computers & Geosciences*, Volume 33, pp. 1301-1315.
- Hewitt, A. & Clark, M., 2013. Initial investigation into the impact of enclosure structures on temperature records. *Met Office Internal Report*.
- Hijmans, R. et al., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, Volume 25, pp. 1965-1978.
- Hofstra, N. et al., 2008. Comparison of six methods for the interpolation of daily, European climate data. *Journal of Geophysical Research*, Volume 13.
- Huband, N. D. S., 1990. Temperature and humidity measurements on automatic weather stations. A comparison of radiation shields.. *Internal Report. Campbell Scientific Ltd. Shepshed, UK*.
- Hubbard, K. G., Lin, X. & Walter-Shea, E. A., 2001. The effectiveness of the ASOS, MMTS, Gill, and CRS air temperature radiation shields. *Journal of Atmospheric and Oceanic Technology*, 18(6), pp. 851-864.

- Hunter, J., Alabri, A. & Van Ingen, C., 2013. Assessing the quality and trustworthiness of citizen science data. *Concurrency Computation Practice and Experience*, 25(4), pp. 454-466.
- Illingworth, S., Muller, C., Graves, R. & Chapman, L., 2014. UK Citizen Rainfall Network: a pilot study. *Weather*, 69(8), pp. 203-207.
- Ingleby, B., Moore, D., Sloan, C. & Dunn, R., 2013. Evolution and Accuracy of Surface Humidity Reports. *Journal of Atmospheric and Oceanic Technology*, Volume 30, pp. 2025-2043.
- Jarvis, C. & Stuart, N., 2001. A Comparison among Strategies for Interpolating Maximum and Minimum Daily Air Temperatures. Part I: The Selection of “Guiding” Topographic and Land Cover Variables. *Journal of Applied Meteorology*, Volume 40, pp. 1060-1074.
- Jeffrey, S., Carter, J., Moodie, K. & Beswick, A., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, Volume 16, pp. 309-330.
- Jenkins, G., 2014. A comparison between two types of widely used weather stations. *Weather*, Volume 69, pp. 105-110.
- Jenkins, G., 2015. Simple investigations of local microclimates using an affordable USB temperature logger. *Weather*, 70(3), pp. 85-88.
- Jones, P. & Lister, D., 2009. The urban heat island in Central London and urban-related warming trends in Central London since 1900. *Weather*, 64(12), pp. 323-327.
- Joo, S., J. E. & Marriott, R., 2013. The Impact of MetOp and Other Satellite Data within the Met Office Global NWP System Using an Adjoint-Based Sensitivity Method. *Monthly Weather Review*, 141(10), pp. 3331-3342.
- Kumamoto, M., Otsuka, M., Sakai, T. & Aoyagi, T., 2012. Field experiment on the effects of a nearby asphalt road on temperature measurement. *WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation*.
- Lacombe, M., Bousri, D., Leroy, M. & Mezred, M., 2011. *WMO Field Intercomparison of thermomemter screens/shields and humidity measuring intruments*, s.l.: WMO instruments and observing methods report No. 106.
- Lefèvre, M. et al., 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmos. Meas. Tech*, Volume 6, pp. 2403-2418.

- Lillesand, T., Kiefer, R. & Chipman, J., 2008. *Remote Sensing and Image Interpretation*. 6th ed. s.l.:John Wiley & Sons.
- Lorenc, A. et al., 2000. The Met. Office global three-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 126(570), pp. 2991-3012.
- Lussana, C., Ubaldi, F. & Salvati, M., 2010. A spatial consistency test for surface observations from mesoscale meteorological networks. *Quarterly Journal of the Royal Meteorological Society*, Volume 136, pp. 1075-1088.
- Mander, N., 2012. An Introduction to the QA Lab. *Met Office Internal Report for Observing System and Data Appreciation Course*.
- Morris, C. & Endfield, G., 2012. Exploring contemporary amateur meteorology through an historical lens. *Weather*, 67(1), pp. 4-8.
- Morton, D. et al., 2011. *Final report for LCM2007 – the new UK land cover map. CS Technical Report No 11/07 NERC/Centre for Ecology & Hydrology 112pp. (CEH project number: C03259)*
- Moseley, S., 2011. From Observations to Forecasts – Part 12 : Getting the most out of model data. *Weather*, 66(10), pp. 272-276.
- Muller, C. et al., 2013. Toward a Standardized Metadata Protocol For Urban Meteorological Networks. *Bulletin of the American Meteorological Society*, Volume 94, pp. 1161-1185.
- Muller, C. et al., 2015. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, Volume Early View Online Version.
- Muller, C. L., 2013. Mapping snow depth across the West Midlands using social media-generated data. *Weather*, Volume 68, p. 82.
- Nabney, I., 2002. *NETLAB: Algorithms for Pattern Recognition*. 1st ed. s.l.:Springer Science & Business Media.
- Nov, O., Arazy, O. & Anderson, D., 2014. Scientists@Home: What drives the quantity and quality of online citizen science participation?. *PloS ONE*, 9(4).
- Oke, T., 2004. *Initial Guidance to Obtain Representative Meteorological Observations at Urban Sites. Instruments and Methods of Observation Program, IOM Report No. 81, WMO/TD 1250*, Geneva: World Meteorological Organization.

- Oke, T. R., 2004. Siting and exposure of meteorological instruments at urban sites. *27th NATO/CCMS International Technical Meeting on Air Pollution Modelling and its Application Banff, Canada, 25-29 October 2004*.
- Orlowsky, B. & Seneviratne, S., 2014. On the spatial representativeness of temporal dynamics at European weather stations. *International Journal of Climatology*, 34(10), pp. 3154-3160.
- Overton, A. K., 2007. *A guide to the siting, exposure and calibration of automatic weather stations for synoptic and climatological observations*. [Online] Available at: <http://myweb.tiscali.co.uk/awsguide> [Accessed 10 October 2013].
- Reda, I. & Andreas, A., 2004. Solar position algorithm for solar radiation applications. *Solar Energy*, Volume 76, pp. 577-589.
- Reed, M., 2011. An investigation into the effect of the synoptic weather on sea breezes at Whitsand Bay, Cornwall. *Weather*, 66(4), pp. 94-97.
- Reno, M. J., Hansen, C. W. & Stein, J. S., 2012. Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis. *Sandia Report*, Volume SAND2012-2389.
- Rigol, J., Jarvis, C. & Stuart, N., 2001. Artificial neural networks as a tool for spatial interpolation. *International Journal of Geophysical Information Science*, 15(4), pp. 323-343.
- Rigollier, C., Lefevre, M. & Wald, L., 2004. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. *Solar Energy*, Volume 77, pp. 159-169.
- Robledo, L. & Soler, A., 2000. Luminous efficacy of global solar radiation for clear skies. *Energy Conversion and Management*, Volume 41, pp. 1769-1779.
- Sevruk, B. & Nespor, V., 1994. The Effect of Dimensions and Shape of Precipitation Gauges on the Wind-Induced Error. *Global Precipitations and Climate Change, NATO ASI Series*, Volume 26, pp. 231-246.
- Sheridan, P., Smith, S., Brown, A. & Vosper, S., 2010. A simple height-based correction for temperature downscaling in complex terrain. *Meteorological Applications*, Volume 17, pp. 329-339.
- Simpson, J., Mansfield, D. & Milford, J., 1977. Inland penetration of sea-breeze fronts. *Quarterly Journal of the Royal Meteorological Society*, 103(435), pp. 47-76.

- Smith, C. et al., 2011. Fine-scale spatial temperature patterns across a UK conurbation. *Climatic Change*, Volume 109, pp. 269-286.
- Steenneveld, G. J. et al., 2011. Quantifying urban heat island effects and human comfort for cities of variable size and urban morphology in the Netherlands. *J. Geophys. Res.*, Volume 116, p. D20129.
- Stewart, I., Oke, T. & Krayenhoff, E., 2014. Evaluation of the 'local climate zone' scheme using temperature observations and model simulations. *International Journal of Climatology*, 34(4), pp. 1062-1080.
- Strangeways, I., 2003. *Measuring the Natural Environment*. 2 ed. Cambridge, UK: Cambridge University Press.
- Strangeways, I., 2007. *Precipitation Theory, Measurement and Distribution*. 1st ed. Cambridge, UK: Cambridge University Press.
- Tang, Y., Lean, H. & Bornemann, J., 2013. The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteorological Applications*, Volume 20, pp. 417-426.
- Thunis, P. & Bornstein, R., 1996. Hierarchy of Mesoscale Flow Assumptions and Equations. *Journal of the Atmospheric Sciences*, 53(3), pp. 380-397.
- Tomlinson, C., Chapman, L., Thornes, J. & Baker, C., 2012. Derivation of Birmingham's summer surface urban heat island from MODIS satellite images. *International Journal of Climatology*, Volume 32, pp. 214-224.
- Visscher, G. J. W. & Kornet, J. G., 1994. Longterm tests of capacitive humidity sensors. *Meas. Sci. Technol.*, Volume 5, pp. 1294-1302.
- Vosper, S. et al., 2014. Cold-pool formation in a narrow valley. *Quarterly Journal of the Royal Meteorological Society*, 140(679), pp. 699-714.
- Waller, J. et al., 2013. Representativity error for temperature and humidity using the Met Office high-resolution model. *Quarterly Journal of the Royal Meteorological Society*, 140(681), pp. 1189-1197.
- Weather, 2012. Weather News. *Weather*, 67(1), p. 2.
- Wheeler, E. F., Zajackowski, J. L. & Graves, R. E., 2003. Effect of solar shielding on portable datalogger temperature readings. *Applied Engineering in Agriculture*, 19(4), pp. 473-481.

- Wilby, R. L., Jones, P. D. & Lister, D. H., 2011. Decadal variations in the nocturnal heat island of London. *Weather*, Volume 66, pp. 59-64.
- Williams, M. et al., 2011. Automatic processing, quality assurance and serving of real-time weather data. *Comput. Geosci.*, Volume 37, pp. 353-362.
- WMO, 2010. *Guide to Meteorological Instruments and Methods of Observation*. WMO-No. 8 (2008 edition, Updated in 2010), Geneva, Switzerland: World Meteorological Organization.
- Wolters, D. & Brandsma, T., 2012. Estimating the Urban Heat Island in residential areas in the Netherlands using observations by weather amateurs. *J. Appl. Meteorol. Climatol.*, Volume 51, pp. 711-721.
- Xia, Y., Winterhalter, M. & Fabian, P., 2000. Interpolation of Daily Global Solar Radiation with Thin Plate Smoothing Splines. *Theoretical and Applied Climatology*, Volume 66, pp. 109-115.

8. Appendix

8.1. Equation notation

This thesis contains several equations – most notably in Section 4.3 and Section 5.6. These equations adhere to the following notation for readability and continuity. A list of all the symbols used in these sections is also included below.

μ & v – For variables modelled as Gaussian distributions; μ is used to denote the mean, and v the variance. v was chosen over the more common notation σ^2 to improve readability when superscripts are used, as detailed next.

Superscripts – Superscripts are used when a symbol is employed multiple times but for different variables or parameters. For example μ and v are used multiple times to represent the Gaussian distribution of several variables. As an example the superscript *Cal* in the notation μ^{Cal} , denotes that this is the mean of the Gaussian distribution that represents the Calibration bias.

Subscripts – Subscript notation is used to indicate that many processes are performed at every timestep, for every station, and sometimes for every design class. For example, the probability that a station belongs to a given design type is notated as $p_{d,s,t}$, because it is calculated for every design class, d , for every station, s , and at each timestep, t .

Bayesian Updating – Many of the models introduced later iterate through time taking initial information and updating it with current estimates to produce a posterior distribution in a Bayesian manner (Bishop, 2007 p17). Our initial distribution is usually the posterior from the previous timestep forecast forward to the current timestep for which subscript $t-1$ represents the previous timestep and tilde, \sim , indicates the ‘prior’ after it has been forecast forward. New data is given the hat notation, $\hat{\cdot}$. Therefore, using the mean of the calibration bias as an example, the posterior from the previous timestep, $\mu_{s,t-1}^{Cal}$, is forecast forward to give the ‘prior’, $\tilde{\mu}_{s,t}^{Cal}$, which is updated with the arrival of new data, $\hat{\mu}_{s,t}^{Cal}$, to give the Posterior, $\mu_{s,t}^{Cal}$.

Interpolation Model	
Symbol	Description
y	‘True’ air temperature at a given weather station location. What the interpolation model aims to predict.
β	Regression coefficient parameter vector. Also shown in its transpose form β^\top .
X	Design matrix. Contains basis functions. Dimensions: $m \times n$.
ϵ	Predicted error of the interpolation model. Represented by a Gaussian distribution of the form: $\epsilon \sim \mathcal{N}(0, v_\epsilon)$
m	Number of weather stations.
n	Number of basis functions (see Section 4.4).
$\sim \mathcal{N}(0, \sigma^2)$	Notation used to indicate a Gaussian (Normal) distribution, shown in this example with mean 0, and variance σ^2 .
$\sim \mathcal{IG}(a, b)$	Notation used to indicate an Inverse Gamma distribution, with shape parameter a , and scale parameter b .
v_ϵ	Model uncertainty. Represented by an Inverse Gamma distribution with the form: $v_\epsilon \sim \mathcal{IG}(a, b)$
a	Generic form of the shape parameter of the Inverse Gamma distribution used to represent model uncertainty, v_ϵ .
a_{t-1}	Shape parameter of the Inverse Gamma distribution representing model uncertainty, v_ϵ , from the previous timestep ($t-1$). A scalar.
\tilde{a}_t	<i>Prior</i> estimate of the shape parameter of the Inverse Gamma distribution representing model uncertainty, v_ϵ . A scalar.
a_t	<i>Posterior</i> estimate of the shape parameter of the Inverse Gamma distribution representing model uncertainty, v_ϵ . A scalar.
b	Generic form of the scale parameter of the Inverse Gamma distribution used to represent model uncertainty, v_ϵ .
b_{t-1}	Scale parameter of the Inverse Gamma distribution representing model uncertainty, v_ϵ , from the previous timestep ($t-1$). A scalar.
\tilde{b}_t	<i>Prior</i> estimate of the scale parameter of the Inverse Gamma

	distribution representing model uncertainty, v_ϵ . A scalar.
b_t	<i>Posterior</i> estimate of the scale parameter of the Inverse Gamma distribution representing model uncertainty, v_ϵ . A scalar.
μ_β	Generic notation for regression coefficient mean term.
$\mu_{\beta,t-1}$	Estimate of the mean of every regression coefficient as learnt at the last timestep ($t-1$). Dimensions: $n \times 1$.
$\tilde{\mu}_{\beta,t}$	<i>Prior</i> estimate of the mean of every regression coefficient. Dimensions: $n \times 1$.
$\mu_{\beta,t}$	<i>Posterior</i> estimate of the mean of every regression coefficient. The result of updating $\tilde{\mu}_{\beta,t}$ with new data at the given timestep. Dimensions: $n \times 1$.
Σ_β	Generic notation for regression coefficient covariance matrix.
$\Sigma_{\beta,t-1}$	Regression coefficient covariance matrix from the previous timestep ($t-1$). Dimensions: $n \times n$.
$\tilde{\Sigma}_{\beta,t}$	<i>Prior</i> estimate of coefficient covariance matrix. Dimensions: $n \times n$.
$\tilde{\Sigma}_{\beta,t}^{-1}$	Regression coefficient precision matrix. The inverse of the square matrix $\tilde{\Sigma}_{\beta,t}$. Dimensions: $n \times n$.
$\Sigma_{\beta,t}$	<i>Posterior</i> regression coefficient covariance matrix. The result of updating $\tilde{\Sigma}_{\beta,t}$ with new data at the given timestep. Dimensions: $n \times n$.
$\Sigma_{\beta,t}^{-1}$	<i>Posterior</i> Regression coefficient precision matrix. Inverse of the square matrix $\Sigma_{\beta,t}$. Dimensions: $n \times n$.
$\Sigma_{\beta,t=0}$	Regression coefficient covariance matrix as initialised at the first timestep. Dimensions: $n \times n$.
δt	The time between the last timestep ($t-1$) and the current timestep (t). Must be in the same units as γ_β . In this study $\delta t = 3$ hours.
γ_β	Forgetting rate parameter for the regression coefficients. In this study $\gamma_\beta = 24$ hours.
Γ	Regularisation term, a diagonal matrix with the value of 0.000001 along the diagonal used to ensure numerical stability. Dimensions:

	$n \times n$.
t	Target temperature observations, used to train the model. Dimensions: $m \times 1$. Not to be confused with the subscript t used to indicate a timestep.
$\mu_{y,t}$	Mean estimate of ‘true’ temperature, y , at the target location at timestep, t .
$\Sigma_{y,t}$	Variance of the ‘true’ temperature estimate at the target location, at timestep, t .
Radiation Specific Interpolation	
l_{Rad}	Global radiation observations that have been weighted temporally over a 60 minute window using an exponential weighting before applying a log transformation.
w_{Rad}	The point minute resolution global radiation observations over the proceeding hour, weighted exponentially.
x	Number of minutes since the time of the temperature observation.
λ	The exponential decay constant.
Bias Model	
Symbol	Description
$\mu_{s,t}^{Rad}$	Mean estimate of the radiation-induced temperature bias for each station, s , at a timestep, t .
$v_{s,t}^{Rad}$	Variance estimate of the radiation-induced temperature bias for each station, s , at a timestep, t .
$\mu_{d,s,t}^{prb}$	The mean radiation-induced temperature bias as estimated by a given design class, d , at a given station location, s , at the timestep, t .
$v_{d,s,t}^{prb}$	The variance term of the radiation-induced temperature bias as estimated by a given design class, d , at a given station location, s , at the timestep, t .
$\mu_{s,t}^{orb}$	A mean estimate of the observed radiation-induced temperature bias at given station location, s , at the timestep, t .
$v_{s,t}^{orb}$	The variance of the observed radiation-induced temperature bias

	estimate at given station location, s , at the timestep, t .
$p_{d,s,t-1}$	The probability that a given station, s , belongs to each design class, d , as estimate at the previous timestep, $t-1$. Range: 0 – 1.
$\tilde{p}_{d,s,t}$	The <i>prior</i> probability that a given station, s , belongs to each design class, d , at the timestep, t . Range: 0 – 1.
$\hat{p}_{d,s,t}$	The probability that a given station, s , belongs to each design class, d , based upon the new data available at this timestep, t . Range: 0 – 1.
$p_{d,s,t}$	The <i>posterior</i> probability that a given station, s , belongs to each design class, d , at the timestep, t . Range: 0 – 1.
p_d^{eq}	Probability of belong to each design class, d , when each has an equal probability.
$\mu_{s,t-1}^{Cal}$	Mean estimate of the temperature calibration bias, for a given station, s , as estimated at the previous timestep, $t-1$.
$v_{s,t-1}^{Cal}$	Variance of the temperature calibration bias estimate, for a given station, s , as estimated at the previous timestep, $t-1$.
$\tilde{\mu}_{s,t}^{Cal}$	<i>Prior</i> mean estimate of the temperature calibration bias, for a given station, s , at the timestep, t .
$\tilde{v}_{s,t}^{Cal}$	<i>Prior</i> variance of the temperature calibration bias estimate, for a given station, s , at the timestep, t .
$v_{s,t=0}^{Cal}$	Variance of the temperature calibration bias estimate, for a given station, s , as estimated at the initial timestep, $t=0$.
$\hat{\mu}_{s,t}^{Cal}$	The mean estimate of the temperature calibration bias for a given station, s , as estimated from new data available at the current timestep, t .
$\hat{v}_{s,t}^{Cal}$	The variance of the temperature calibration bias estimate for a given station, s , as estimated from new data available at the current timestep, t .
$\mu_{s,t}^{Cal}$	<i>Posterior</i> mean estimate of the temperature calibration bias, for a given station, s , at the timestep, t .
$v_{s,t}^{Cal}$	<i>Posterior</i> variance of the temperature calibration bias estimate, for a given station, s , at the timestep, t . Used as a representativity

	term.
$\delta t_{s,t}$	The timestep between the current observation at timestep, t , by a station, s , and it's last observation. Units should be the same as the forgetting and learning rate parameters γ and α .
γ^{cal}	Forgetting rate parameter for the calibration bias. In this study $\gamma^{cal} = 50$ days.
γ^{dp}	Forgetting rate parameter for the design class membership probabilities. In this study $\gamma^{dp} = 24$ hours.
α^{cal}	Learning rate parameter for the calibration bias. In this study $\alpha^{cal} = 20$ days.
α^{dp}	Learning rate parameter for the design class membership probabilities. In this study $\alpha^{dp} = 6$ hours.
$S_{d,s,t}$	Scaling factor for given design class, d , for a given station, s , at the timestep, t .
$\mu_{s,t}^{CWSu}$	Mean uncorrected CWS observation for a given station, s , at the timestep, t .
$v_{s,t}^{CWSu}$	Variance of uncorrected CWS observation for a given station, s , at the timestep, t . In this study $v_{s,t}^{CWSu} = 0.2$ °C.
$\mu_{s,t}^{CWSc}$	Mean corrected CWS observation for a given station, s , at the timestep, t .
$v_{s,t}^{CWSu}$	Variance of corrected CWS observation for a given station, s , at the timestep, t .
$\mu_{s,t}^{IMMS}$	Mean estimate of temperature at station location, s , at the timestep, t , estimated by interpolating MMS temperature observations to the target station location. Calculated in Equation (10).
$v_{s,t}^{IMMS}$	Variance of the temperature estimate at station location, s , at the timestep, t , estimated by interpolate model when interpolating MMS temperature observations to the target station location. Variances are taken off the diagonal of $\Sigma_{y,t}$ as calculated in Equation (11) .

8.2. Weather Underground station types

Weather station manufacturers and models used to automatically upload data to Weather Underground in February 2012. The raw list of stations was accessed from:

<http://www.wunderground.com/weatherstation/ListStations.asp?selectedCountry=United+Kingdom>

The raw list of stations were organised into the following models of station, whilst accounting for differences in spelling.

		Count	% of known
WH1080	WH1080 (including ones with solar powered transmitter)	335	29.67
Davis	Pre Vantage Pro Davis systems	14	1.24
	Davis Vantage Vue	73	6.47
	Davis Vantage Pro	46	4.07
	Davis Vantage Pro Plus	13	1.15
	Davis Vantage Pro 2	199	17.63
	Davis Vantage Pro 2 Plus (w/o FARS)	37	3.28
	Davis Vantage Pro 2 FARS (w/o Plus)	22	1.95
	Davis Vantage Pro 2 Plus FARS	4	0.35
	Unknown Davis Vantage	12	1.06
La Crosse	La Crosse with wind fan, temp in box, rect rain guage	52	4.61
	La Crosse with wind cups, temp in box, rect rain guage	38	3.37
	Lacrosse unknown or less common	26	2.30
Oregon Scientific	Has little transmitter for each, temp probe WMR928 & similar	81	7.17
	Oregon Scientific similar to the WMR200, wind cups, unusual temp box	69	6.11
	Oregon Scientific WMR100 (only)	35	3.10
	Other OS systems e.g. WMR88 and Oregon Unknown or Less Common	22	1.95
Others	One-Wire Systems	24	2.13
	Nexus and Ventus and Irox	15	1.33
	Others (Including the Rainwise MKIII and WS-2000)	12	1.06
	unknown	224	
	total known	1129	
	total	1353	
	Percentage unknown	16.56	

8.3. WOW station types

Count made on 2nd Dec 2013. Total of 1116 CWS uploading to WOW globally. 770 had written some textual metadata about their site. Written in the *Site Description* and *Additional Information* sections. From this metadata the station type could be derived, using a well-trained automatic keyword search, from just 485 stations.

The following list shows the percentage of these 485 stations made up by each model of station:

- **Fine Offset WH1080** (includes rebranded versions) - Accounts for **31.5%** of the 485 stations.
- **Davis Vantage Pro2** (without fan aspiration) - **28.5%**

- **Davis Vantage Vue** - 10.3%
- **Davis Vantage Pro** - 6.8%
- **Davis Vantage Pro2** (with fan aspiration) - 5.2%
- **La Crosse WS2300** (and similar looking models) - 4.7%
- **Oregon Scientific WMR200** (and similar) - 3.1%
- **1-Wire systems** (such as Dallas and AAG) - 2.7%
- 2.5% mentioned owning a **Stevenson Screen**
- **Oregon Scientific WMR100** (and similar) - 2.3%
- **Oregon Scientific WMR968** (and similar) - 2.1%
- **Oregon Scientific WMR88** (and similar) - 1.9%
- **La Crosse WS3600** (and similar) - 0.8%
- **TFA 35-1095** (and similar) - 0.8%

Were this list to continue it would simply show station models for which there are only 1 or 2 counts of. These include other La Crosse, Oregon Scientific and Davis models as well as some more advanced kits by manufacturers such as Vaisala, Skye, Peet and Columbia weather systems.

The list often states 'and similar looking models'. This is as for companies such as Oregon Scientific and La Crosse many of their stations look very similar, but have different model numbers. For example the OS WMR200 and WMR180 are very similar so both have been classed as the WMR200.

During this process around 30 different rebranded versions of the Fine Offset WH1080 were counted. Including Watson, Ambient Weather, Tycon, Nevada, Maplin, Weathereye, ClimeMet, Jenkinsbird, Elecsa, Weatherwise and Aercus.

8.4. WOW count code

Code written in the language *Ruby*, and run on a daily Cron Job, to count the number of stations uploading to WOW each day.

```

#-----
# Aim
#-----
# The end goal of this code is to produce a JSON that highcharts can use to plot up the number
# of WOW stations over time.
# It works out the total number in the UK and Ireland, and also globally.
# From a given start date up to the present it finds the count for a given hour every day.
# Once it has run intially this code is then meant to be called every day on the server using a cron job.
# It will then search for any day's it's missing (i.e. the present new day) and scrape the totals for them.

#-----
# Dependencies
#-----
require "open-uri"
require "json"
require "nokogiri"
require "csv"
require "time"
require "pp"

# Specify the hour of the day you'd like to count the number of stations at.
wowHour = "12" # If single digit give padding zero
# With this fixed we should at least be counting the station at the same time of the day every day.
# Want this hour to be at least an hour before the scrape time, just in case all the present
# data isn't up yet.

# Select your start and end dates (Good start date is 2011-01-01)
# Must set the timezone of the start date otherwise the server will set it to it's own timezone
sd = Time.parse("2011-01-01 #{wowHour}:00:00 -0000")
ed = Time.parse("2014-01-01 #{wowHour}:00:00 -0000")
ed = Time.now.utc
# Note how we use the Time class not the Date class as when we used .to_i Later to give the unix epoch time
# this command only works on the Time class

# Set where and what name to save the JSON you create as
saveName = "/home1/weathes4/public_html/wow/wowCountJSONv2.json"

# If you've already created your JSON file then Load it in as a hash here.
if File.exists? saveName

  puts "Previous JSON File Exists, loading."
  cHash = JSON.parse(IO.read(saveName))

else

  puts "No previous JSON found"
  # If you haven't already the hash to hold the counts
  cHash = {}

end

# Begin by setting your LoopTime as the start date, LoopTime is of class Time.
loopTime = sd

while loopTime <= ed - (60*100)
  # I've add this subtraction of 100 minutes to ensure the code won't run for the present day if you're scraping to
  # close to the time you're counting
  begin # For rescue

    # Convert the LoopTime into the format you'll use as the key name
    keyDate = loopTime.strftime('%Y-%m-%d %H:%M:%S')

    # First Lets check we haven't already got counts for this time, for both the UK and Globally.
    # If so we don't need to bother scraping them again.
    unless cHash.include?(keyDate)

      theday = loopTime.strftime('%d') # Use strftime to give preceeding zeros, rather than .day
      themonth = loopTime.strftime('%m')
      theyear = loopTime.strftime('%Y')
      thehour = loopTime.strftime('%H')

      # Do you want to return met office stations too? 'on' or 'off'
      wantMMS = 'off'

      theURI = "http://wow.metoffice.gov.uk/ajax/home/map?timePointSlider=0&timePointPicker=-1&northLat=62&southLat=48&eastLon=11&westLon=-:"
      # Changing the lat and lon bounding box values don't seem to change how many stations are returned
      # The timePointSlider is set to 0, this is what it is set to when you select a specific time by dragging
      # the slider all the way to the Left on the WOW user interface. So it should give you the values for the
      # Exact time your after

      #puts theURI

      uri = open(theURI)
      sites = JSON.parse(uri.read)

      #puts JSON.pretty_generate(sites)

      # You've began looking at a new time so reset the counts
      countLCM = 0
      countUKIRE = 0
      countGlobal = 0

      # Loop through each site
      @sites = sites["r"].collect do |site|

        # jtwr stations are the ones we want, stands for those displayed on the map

```

```

if site["mt"] == "jtwr"

  id = site["msi"]
  slat = site["mla"]
  slon = site["mlo"]

  # Add the station to global count
  countGlobal +=1

  # Convert Lat Lon strings to floats
  latf = slat.to_f
  lonf = slon.to_f

  # But is it located within the UKIRE and Lcm2007 extents?
  if (latf > 49.8) && (latf < 61.0) && (lonf > -10.8) && (lonf < 1.78)
    #puts "Station is within British Isles Bounding box"

    if (latf < 51.9) && (latf > 49.0) && (lonf > 1.5) && (lonf < 6.0)
      #puts " However the station is located in France"

    elsif (latf < 50.4) && (latf > 49.0) && (lonf > -3.3) && (lonf < 6.0)
      #puts " However the station is located in France"

    elsif (latf > 55.0) && (latf < 55.5) && (lonf > -12 ) && (lonf < -5.9)
      #puts " However the station is located in Ireland"
      countUKIRE +=1

    elsif (latf > 53.9) && (latf < 55.1) && (lonf > -12 ) && (lonf < -5.3)
      #puts " However the station is located in Ireland"
      countUKIRE +=1

    elsif (latf > 51.3) && (latf < 55.0) && (lonf > -12 ) && (lonf < -5.5)
      #puts " However the station is located in Ireland"
      countUKIRE +=1

    elsif (latf > 53.9) && (latf < 54.5) && (lonf > -5 ) && (lonf < -4 )
      #puts " However the station is located on the Isle of Man"
      countUKIRE +=1

    elsif (latf < 59.6) && (latf > 59.4) && (lonf < -1.5) && (lonf > -1.7)
      #puts " However the station is located on the Fair Isle"
      countUKIRE +=1

    else

      countUKIRE +=1
      countLCM += 1

    end

  else

    #puts "Station isn't located within British Isles bounding box"

  end

end #Site jtwr?
end # Each site

puts "***** #{loopTime.strftime('%d %b %Y %H:%M')} *****"
puts "#{countGlobal} Global stations"
puts "#{countUKIRE} UK and Ireland stations"
puts "#{countLCM} LCM2007 stations"
puts " "

# Add this date and the counts to the cHash
cHash[keyDate] = {"global" => countGlobal, "ukire" => countUKIRE, "lcm"=> countLCM}

else
  puts "Already have counts for #{keyDate}"
end # Unless

rescue
  # You might want to remove this rescue once it's at the point of running daily.
  puts "*** Rescuing for #{loopTime} ***"
  puts " "
end #for rescue

# Add a day onto your date
loopTime = loopTime + (60*60*24)

end #each LoopTime

# Save the hash as a json. Note it also sorts by date, just incase we missed a date first time round that
# got added to the end of the hash the second time round.
File.open("#{saveName}", "w") do |f|
  f.write(Hash[cHash.sort].to_json)
end

puts "New JSON file saved"

```

8.5. WOW site ratings scheme

The WOW website encourages its citizen observers to rate the quality of their weather station based upon a series of attributes listed here:

- Exposure
- Measurements of air temperature
- Measurements of rainfall
- Measurements of wind
- Urban Climate Zone Index (UCZ)
- Reporting hours

These attributes then combine to give the station an overall site rating. The breakdown of this rating system is shown below, and has been taken directly from the UK Met Office's WOW website (<http://wow.metoffice.gov.uk/support/siteratings>).

Exposure

- **5: Very open exposure:** no obstructions within 10h or more of temperature or rainfall instruments.
- **4: Open exposure:** most obstructions/heated buildings 5h or from temperature or rainfall instruments, none within 2h.
- **3: Standard exposure:** no significant obstructions or heated buildings within 2h of temperature or rainfall instruments.
- **2: Restricted exposure:** most obstructions/heated buildings >2h from temperature or rainfall instruments, none within 1h.
- **1: Sheltered exposure:** significant obstructions or heated buildings within 1h of temperature or rainfall instruments.
- **0: Very sheltered exposure:** site obstructions or sensor exposure severely limit exposure to sunshine, wind, rainfall.
- **R: Rooftop site:** Rooftop sites for temperature and rainfall sensors should be avoided where possible.
- **T: Traffic site:** equipment sited adjacent to public highway.
- **U:** Exposure unknown or not stated.

Exposure ratings relate to the site of the temperature and rainfall instruments only, which should ideally be at ground level. Sensors for sunshine, wind speed etc are best exposed as freely as possible, and rooftop or mast mountings are usually preferable.

Exposure guidelines are based on a multiple of the height h of the obstruction above the sensor height; the standard is a minimum distance of twice the height (2h). Thus for a raingauge at 30 cm above ground, a building 5 m high should be at least 9.4 m distant (5 m less 0.3 m, x 2), and a 10 m building should be at least 17 m from a thermometer screen (10 m less 1.5 m, x2)

Measurements of air temperature

- **A:** Standard instruments in Stevenson Screen, calibration within last 10 yr, site exposure minimum rating = 3.
- **B:** Standard instruments in Stevenson Screen or manufacturer supplied AWS radiation screen, calibration within last 10 yr, site exposure = 2 or 3.
- **C:** Standard instruments in Stevenson Screen or manufacturer supplied AWS radiation screen, site exposure 1 or less.
- **D:** Non-standard instruments and/or no or non-standard radiation screen and/or sheltered site, site exposure 1 or less.
- **U:** Instruments unknown or not stated.
- **0:** No air temperature measurements made at this site.

Measurements of rainfall

- **A:** Standard "five inch" manually-read raingauge or calibrated tipping-bucket raingauge, at standard height above ground (30 cm), site exposure minimum = 3.
- **B:** Standard "five inch" manually-read raingauge or calibrated tipping-bucket raingauge, the rim mounted at standard height above ground (30 cm), exposure = 2 or 3.
- **C:** Standard "five inch" manually-read raingauge or calibrated tipping-bucket raingauge, the rim mounted at standard height above ground (30 cm), exposure 1 or less.
- **D:** Non-standard raingauge and/or tipping-bucket raingauge, exposure 1 or less.
- **U:** Instruments unknown or not stated.
- **0:** No rainfall measurements made at this site.

STANDARD INSTRUMENTS in this context means: Standard-pattern (Snowdon or Met Office Mk II pattern) "five-inch" copper raingauge, with deep funnel, the rim of the gauge level and mounted at 30 cm above ground level, meeting the minimum exposure requirement of being at least 'twice the height' of the obstacle away from the obstacle.

Measurements of wind

- **A:** Wind sensors calibrated within last 10 years, mounted 10m above the ground on mast or pole, with no obstructions within 100m.
- **B:** Wind sensors mounted above the ground on mast or pole, with no obstructions within 50m.
- **C:** Wind sensors mounted on building or wall.
- **U:** Instruments unknown or not stated.
- **0:** No wind measurements made at this site.

Urban Climate Zone Index (UCZ)

- **1:** Intensely developed urban zone with detached close-set high-rise buildings with cladding, e.g. downtown towers.
- **2:** Intensely developed high density urban with 2 - 5 storey, attached or very close-set buildings often of brick or stone, e.g. old city core.
- **3:** Highly developed, medium density urban with row or detached but close-set houses, stores & apartments e.g. urban housing
- **4:** Highly developed, low density urban with large low buildings & paved parking, e.g. shopping mall, warehouses.
- **5:** Medium development, low density suburban with 1 or 2 storey houses, e.g. suburban housing.
- **6:** Mixed use with large buildings in open landscape, e.g. institutions such as a hospital, university, airport.
- **7:** Semi-rural development with scattered houses in natural or agricultural area, e.g. farms, estates.
- **U:** UCZ unknown or not stated.

UCZ descriptions as defined by the *World Meteorological Organisation (WMO-No.8, 7th Edition)*

Reporting hours

- **A:** Will always aim to provide a weather report at 09:00 GMT. Daily temperature and rainfall values relate to standard 24 hour period morning to morning.
- **B:** Will always aim to provide a weather report between 06:00 and 09:00 GMT. Daily temperature and rainfall values relate to standard 24 hour period morning to morning.
- **C:** Daily temperature and rainfall values relate to the 24 hour period midnight to midnight. This is the default for most automatic weather stations.
- **D:** Air temperature and rainfall terminal hour is other than A, B or C above, or extremes do not relate to 24 hour periods.
- **U:** Reporting hours unknown or not stated.

How site ratings are calculated

Each site is automatically allocated a 'site rating' based on the observing location attributes entries submitted on site registration. The system is based on the quality and exposure of the temperature and rainfall data:

5* = E5, T=A, R=A

4* = E >= 3, T=A, R=A

3* = E >= 3, T=[A,B or C], R=[A,B or C]

2* = E >= 1, T=[Any], R=[Any]

1* = E =0,1,R or U, T=[Any], R=[Any]

(Where E = Exposure, T = Temperature, and R = Rainfall, and each of these are described in **Location Attributes**).

If temperature is measured at a site, but not rainfall, the site rating will be based on the quality and exposure of the temperature data alone. If rainfall is measured at a site, but not temperature, the site rating will be based on the quality and exposure of the rainfall data alone.

*If there is no temperature or rainfall data, the site will be classed as 1**

8.6. Winterbourne No. 2 field study site metadata

Courtesy of the University of Birmingham's Urban Climate Lab (BUCL).

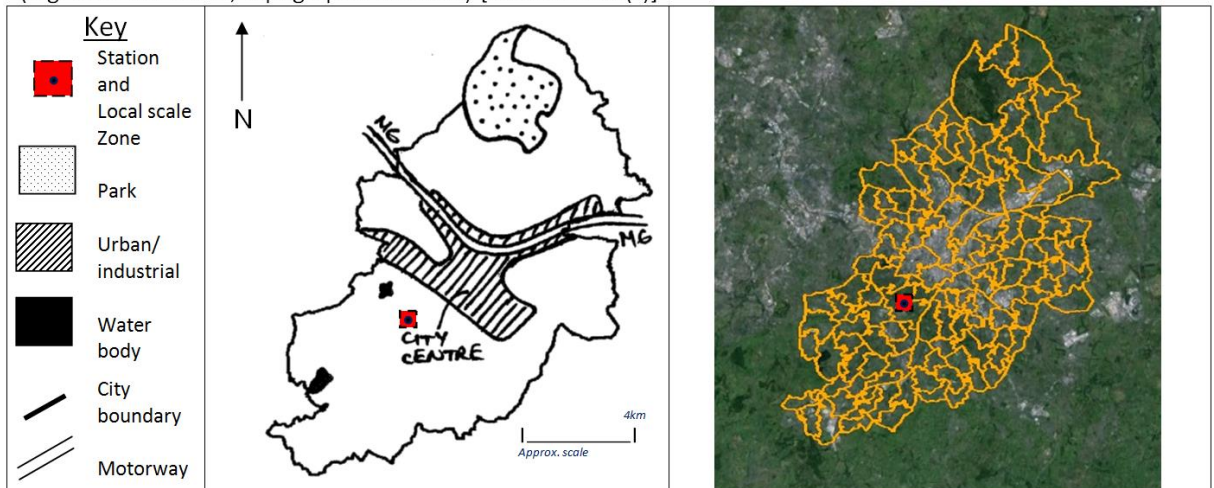
General Station Information

Station Name: Elms Cottage		Commissioned: 06/06/2013	
Station ID: W026	Alias: Uni. Birmingham / Winterbourne 2		Type of Site: University WXT
Elevation 150 m (a.m.s.l.)	Latitude: 52.45638 ° N		Longitude: -1.92767 ° W
Address: Elms Road, Edgbaston, B15 2TT			
Site Contact: Duick Young / Catherine Muller		Phone: 01214149005	
Email: team.bucl@contacts.bham.ac.uk			

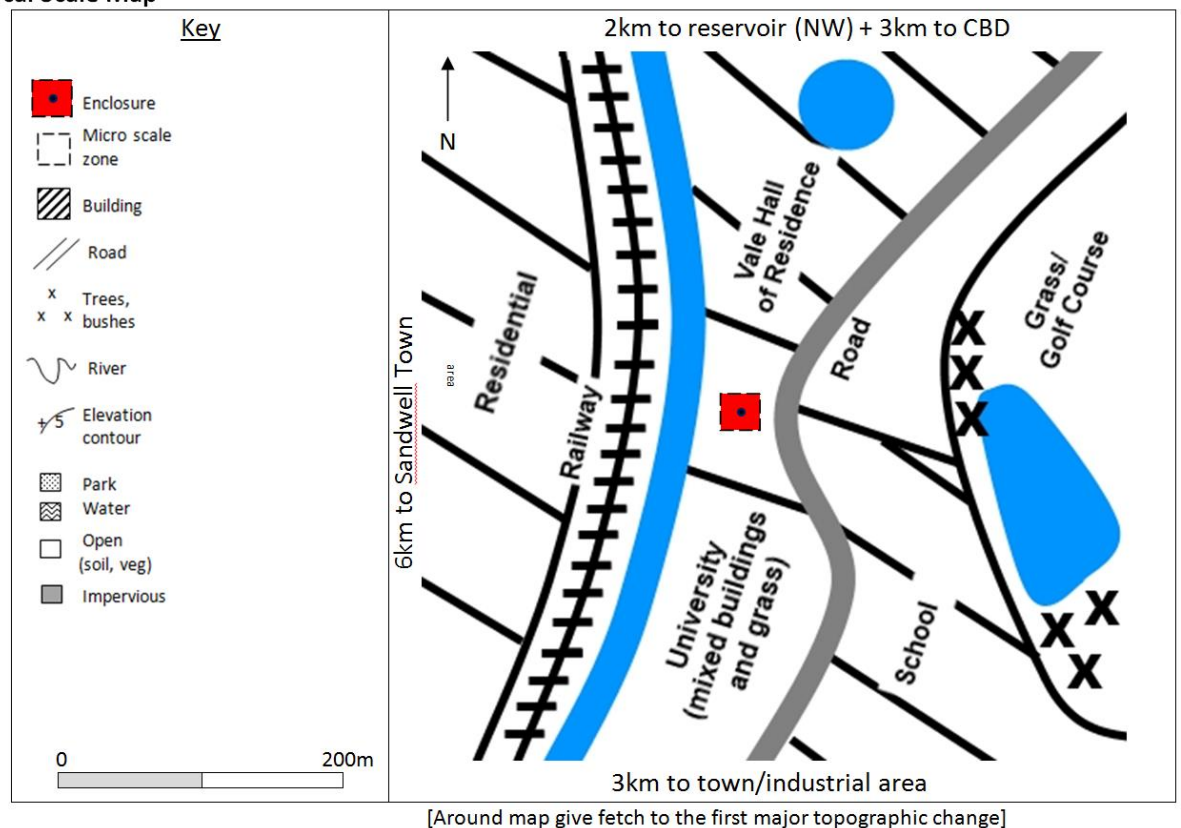
Technical information

Communications network name: University network	Type: Modem transmission to server ("Dave")
SIM Phone number: N/A	SIM no. N/A

City scale map + Aerial photograph(s): showing location of station and local area zone plus important features (e.g. land-use zones, topographic features) [include scale(s)]



Local Scale Map

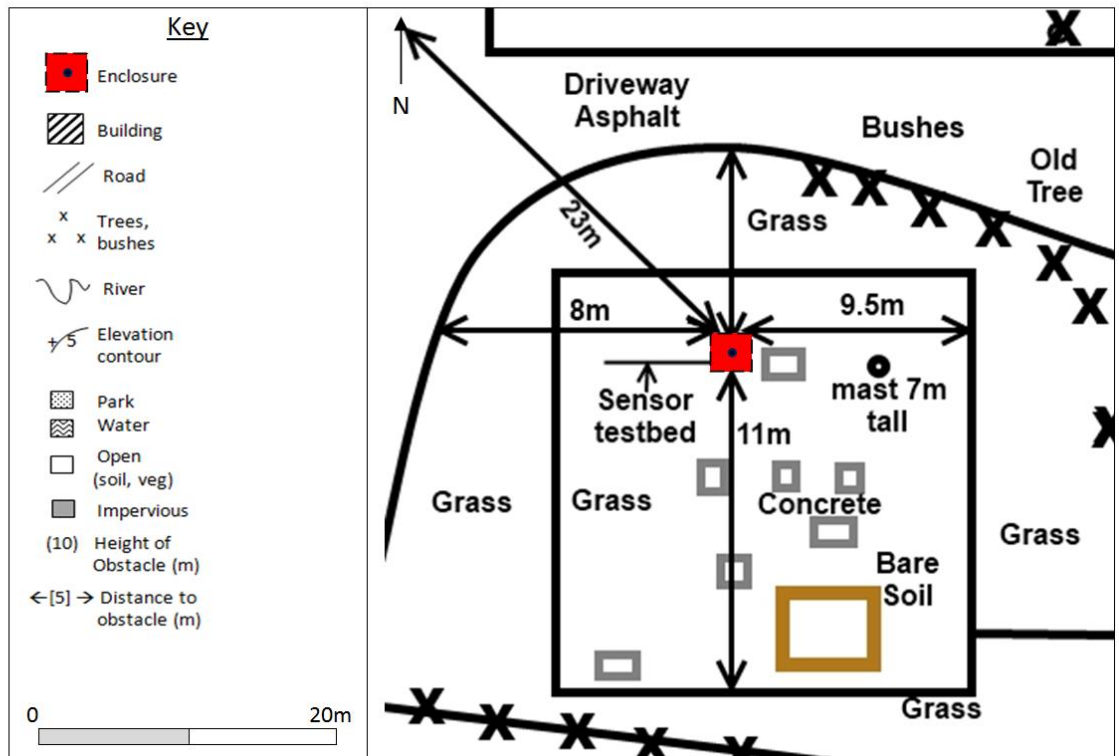


General Information

Local land-classification (e.g. Local Climate Zone [LCZ]): LCZ5 _{6B}									
Dominant Land Use: Residential (semi-detached and large detached houses), University campus, golf course									
Davenport Roughness Class (DRC) upwind of enclosure (500 m):									
6	N	4	E	5	S	6	W		
Land Cover (% in 500 m radius area):									
35% Vegetated		40% Built		15% Impervious		5% Bare		5% Water	
Mean Tree Height:		10 m		Mean Building Height:		3		Storeys	
Lawn/ Garden Irrigation/External water use: golf course; gardens; field (open and sports); Edgbaston pool									
Typical Wall Material: brick				Typical Road Material: asphalt					
Typical Roof Design: gable				Typical Roof Material: slate					
Space heating? No				Space cooling? No					
Traffic Density: light-medium									
Recent Changes/Development since last metadata update (e.g. new residential development):									

Micro Scale (Instrument Exposure)

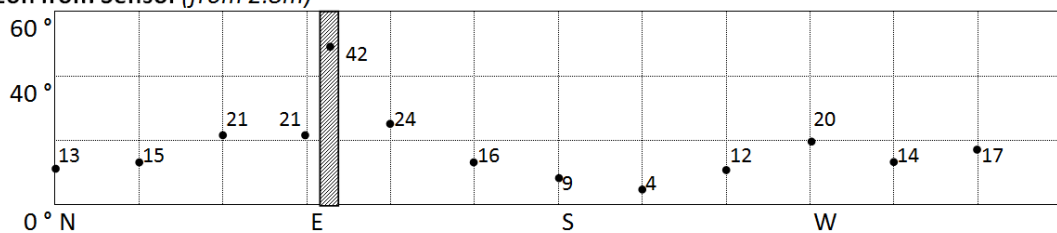
Micro Scale Sketch Map



General Information

Enclosure ID (if applicable):	Elevation: 150+3.3 m (a.m.s.l.)
Latitude: 52.45639 ° N	Longitude: -1.92767 ° W
Mount/enclosure type: custom pole	Mount/enclosure location: field site
Temp/Humidity Sensor Height: 3.3 m	Ventilated shield? No
Surface Cover Below Sensor: grass, concrete, asphalt, bare soil	
Soil/ Material under cover: vegetated compact	Building Materials: brick
Building Types: detached	Building Height: 2 storeys
Tree Type (deciduous/coniferous): both	Tree Height: 20 m
Roof shape: gable	Roof material: clay tile
Site land-classification: LCZ9	
SVF: Winter: 0.89	Summer: (Will be assessed for Summer 2014)
Traffic Density (e.g. none, light, medium, heavy): none	Time: 14:00
Heat / Moisture Vents: none on this side but greenhouse heating to W 40m	

Horizon from Sensor (from 2.8m)



Horizon notes (e.g. notable features):

Tree on edge of site 30m	Conifer by road 30m	Dead tree with ivy 12m	Conifer tree by road 20m	Conifer by road 25m	Large tree by road 50m	Trees by footpath 50m	Tree by IT building 100m	Tree by IT building 100m	Tree by railway and canal 60m	Tree behind greenhouse 50m	Large conifer by canal 50m
--------------------------	---------------------	------------------------	--------------------------	---------------------	------------------------	-----------------------	--------------------------	--------------------------	-------------------------------	----------------------------	----------------------------

Cardinal Photographs (from sensor height and location)

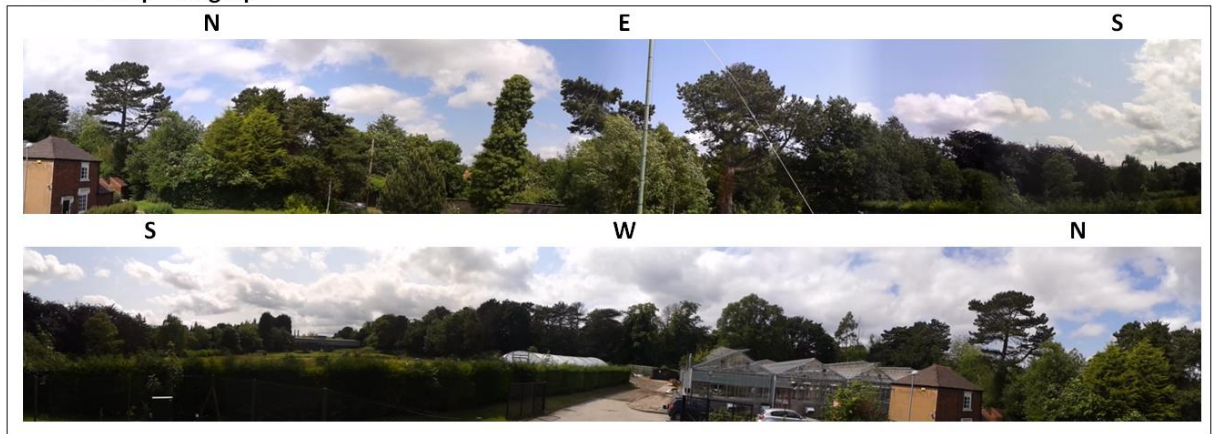




Sky View photograph

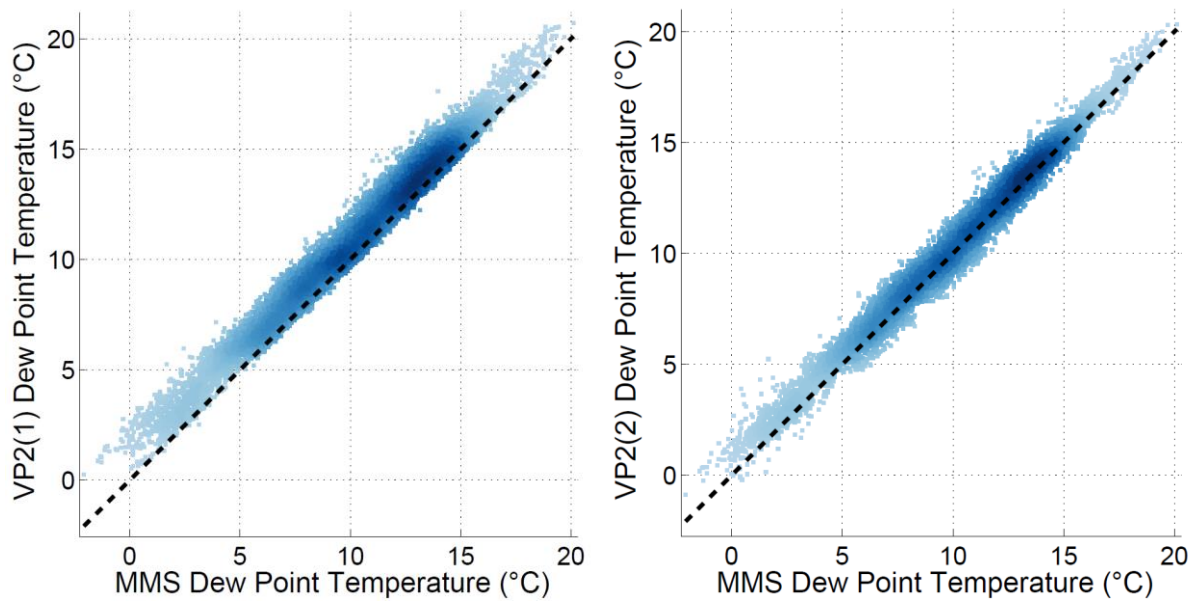


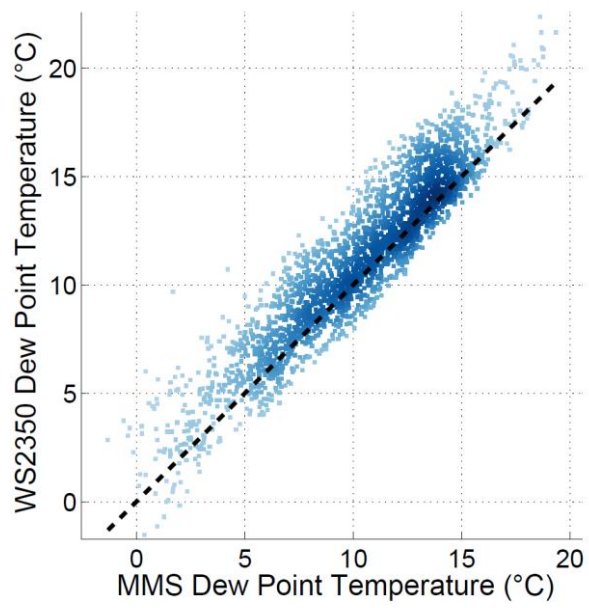
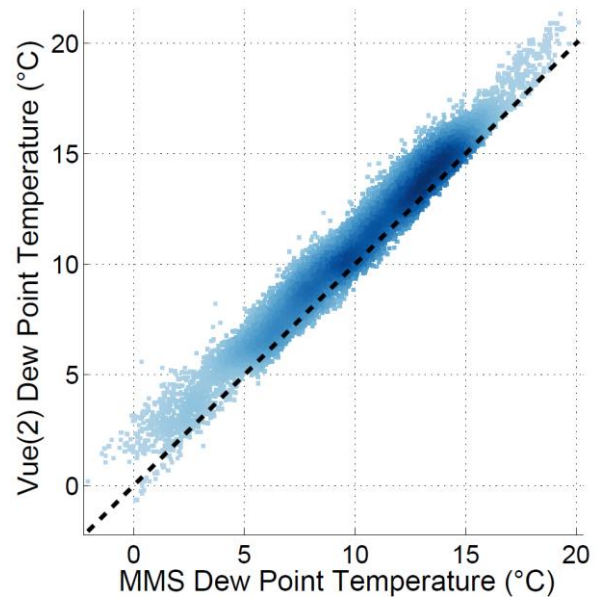
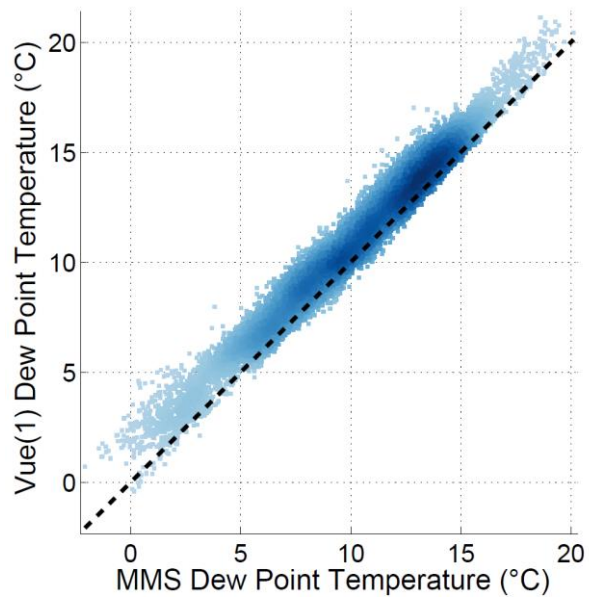
Panoramic photograph



8.7. CWS versus MMS dew point temperature

Show below are the dew point temperature observations for each of the CWS tested in the field study versus the MMS's dew-point temperature. The equivalent plots for the WMR200 and WH1080 are shown in Section 3.2.3. The darker the colour the greater the density of points.





8.8. Rain gauge lab test results

500ml of water was slowly dripped through each rain gauge indoors. This was performed 3 times. By dividing the volume of water by the area of the gauge it is possible to calculate the depth of rain (in mm) that the station's console is expected to display (Overton, 2007). Absolute and percentage differences between expected and measured depth are then calculated. The two VP2s are the only type that could be calibrated first. Screws under the tipping buckets were used to do this. For these two stations they were calibrated as best as possible first before performing this experiment.

Station Nickname	Rain Gauge Dimensions	Gauge area (mm ²)	Tip Resolution/depth (mm)	Tip Volume (ml)	Expected	Reading 1 (mm)	Reading 2 (mm)	Reading 3 (mm)	Average Reading (mm)	Absolute Difference (mm)	Percentage Difference (%)	Correction	Std. Dev. of 3 readings (mm)
VP2(1)	165mm Diameter	21382.465	0.2	4.28	23.38	23.4	22.4	23.0	22.9	-0.5	-1.9	1.0196	0.5
VP2(2)	165mm Diameter	21382.465	0.2	4.28	23.38	22.8	23.4	23.8	23.3	-0.1	-0.2	1.0022	0.5
Vue(1)	122mm diameter	11689.9	0.2	2.34	42.77	39.8	40.0	41.2	40.3	-2.4	-5.7	1.0605	0.8
Vue(2)	122mm diameter	11689.9	0.2	2.34	42.77	37.6	37.2	37.0	37.3	-5.5	-12.9	1.1477	0.3
WMR200	100mm diameter	7854	1.016	7.98	63.66	62.2	62.2	62.2	62.2	-1.5	-2.3	1.0235	0.0
WS2350	55m x 125mm	6875	0.518	3.56	72.73	70.9	71.4	72.0	71.4	-1.3	-1.8	1.0181	0.6
WH1080	51mm x 111mm	5661	0.3	1.70	88.32	100.2	99.6	100.5	100.1	11.8	13.3	0.8824	0.5

